UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Estimación automática de atributos corporales usando redes neuronales convolucionales

Máster Universitario en Ingeniería de Telecomunicación

Autor: CUESTA HIERRO, Helena

Tutor: VERA RODRIGUEZ, Rubén

Ponente: FIÉRREZ AGUILAR, Julián

Dpto. Tecnología Electrónica y de las Comunicaciones

Septiembre, 2020

Estimación automática de atributos corporales usando redes neuronales convolucionales

AUTOR: Helena Cuesta Hierro

TUTOR: Rubén Vera Rodríguez

PONENTE: Julián Fiérrez Aguilar

Biometrics and Data Pattern Analytics Lab

Dpto. Tecnología Electrónica y de las Comunicaciones

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Septiembre 2020

Resumen

Los rasgos biométricos suaves son características que se pueden utilizar para describir a un individuo, pero no sirve para identificar de forma única dado que no son exclusivos de una sola persona. Algunos de estos rasgos son altura, género, piel, color de pelo o ropa.

A diferencia de la biometría primaria como son la huella dactilar, el iris o la voz, no requiere de la cooperación del sujeto para extraer estas características y puede realizarse a distancia mediante cámaras de vigilancia.

La información extraída se puede fusionar con sistemas de reconocimiento biométrico para mejorar el reconocimiento o proporcionar una identificación aproximada, como localizar a un sujeto que ha sido visto anteriormente o que coincide con una descripción.

Estos modelos, aunque no son tan precisos como la biometría tradicional, pueden alcanzar índices aceptables de precisión donde la biometría tradicional no se puede aplicar.

Por este motivo, en este trabajo Fin de Máster se propone el desarrollo e implementación de un clasificador de atributos biométricos suaves en la distancia utilizando redes neuronales convolucionales. Este sistema, puede ser capaz de extraer los atributos de forma automática y realizar una clasificación tras obtener el rendimiento final de la red. Pero para ello, deben haber sido entrenados previamente con una gran base de datos.

Sin embargo, la escasez de imágenes etiquetadas para poder entrenar los diferentes modelos ha supuesto una dificultad que se ha tenido que hacer frente con diferentes técnicas de ampliación de datos.

A lo largo de este trabajo se ha comparado dos redes neuronales comúnmente conocidas en el campo del aprendizaje profundo con la red desarrollada para observar los resultados y determinar la complejidad y parámetros utilizados que mejores resultados obtienen.

En definitiva, el principal objetivo del proyecto ha sido implementar un clasificador de rasgos biométricos suaves, extraídos automáticamente a partir de una imagen de un viandante, tomada en la distancia, con un sistema de redes neuronales convolucionales y comparar los resultados con los obtenidos en las redes AlexNet y GoogLeNet.

Palabras Clave: Redes neuronales convolucionales, clasificador de eventos, rasgos biométricos suaves, atributos corporales.

Abstract

Soft biometrics are characteristics that can be used to describe an individual, but don't work to uniquely identify since they are not exclusive to a single person. These include traits such as height, gender, skin, hair color or clothing.

Unlike traditional biometrics, e.g., fingerprint, iris or voice, it does not require cooperation from the subject to extract these attributes and can be acquired by surveillance cameras.

The extracted information can be fused with biometric recognition systems to improve the overall recognition or provide an approximated identification such as locate an individual that has been seen before or that matches a description.

These models, even though they are not as accurate as traditional biometrics, can achieve acceptable rates of accuracy where traditional biometrics cannot be applied.

For this reason, in this final master Project the development and implementation of a soft biometric attribute classifier in distance using convolutional neural networks is proposed. This system might be able to extract the attributes automatically and perform a classification just after obtaining the accuracy of the network. Before this, they must have been previously trained with a large dataset.

However, the shortage of labeled images to be able to train the different models has been a difficulty that has faced with different data augmentation techniques.

Along this Project, two well-known neural networks have been compared with the developed network to observe the results and determine the best complexity and parameters that gets better results.

In short, the main objective of this Project has been to implement a classifier of soft biometric features, extracted automatically from an image of a pedestrian taken at a distance, with a system of convolutional neural network and to compare the results with those obtained in AlexNet and GoogLeNet networks.

Keywords: Convolutional Neural Networks, Event's Classifier, Soft Biometrics, Corporal Features.

Agradecimientos

Quiero dar las gracias a todas las personas que me han acompañado durante mis estudios universitarios y me han ayudado a hacer este trabajo fin de máster, pero me gustaría mencionar a los más importantes.

En primer lugar, mi familia, mis padres y hermano, que siempre me han apoyado en los momentos más difíciles y me han permitido estudiar lo que siempre quise facilitándome el día a día para que pudiese dedicar el tiempo necesario a los estudios.

A mis amigos, no quiero dejarme a nadie, pero los puedo agrupar principalmente en cuatro grupos: los de la infancia, el grupo 95 de la carrera, los compañeros del máster y los que me acompañan actualmente todos los días en el trabajo. He pasado tantas horas con cada uno de ellos que, sin duda, son mi segunda familia, ¡y menuda familia!

Por último, a mi tutor Rubén, nadie mejor que yo sabe la paciencia que ha tenido. No hay palabras suficientes para agradecer su tiempo y dedicación.

A todos vosotros, muchas gracias.

ÍNDICE GENERAL

1. INTRODUCCIÓN 1

1.1 MOTIVACIÓN 1

1.2 OBJETIVOS 2

1.3 ORGANIZACIÓN DE LA MEMORIA 2

2. ESTADO DEL ARTE 5

2.1 APRENDIZAJE PROFUNDO 5

2.1.1 Tipos de aprendizaje profundo 7

2.1.2 Conjunto de datos o Data Set 8

2.2 REDES NEURONALES ARTIFICIALES 8

2.2.1 Breve historia de las redes neuronales 9

2.2.2 Estructura de una red neuronal 9

2.2.3 Funciones de activación 11

2.2.4 Algoritmo Backpropagation 13

2.2.5 Función de Coste 13

2.2.6 Stochastic Gradient Descent (SGD) 14

2.3 REDES NEURONALES CONVOLUCIONALES 14

2.3.1 Preprocesamiento y feature mapping 15

2.3.2 Aprendizaje por transferencia 16

2.3.3 Entrenamiento de la red 17

2.4 RECONOCIMIENTO BIOMÉTRICO 17

2.4.1 Ventajas de la biometría 18

2.4.2 Partes de un sistema biométrico 19

2.4.3 Modos de operación 20

2.5 RASGOS BIOMÉTRICOS 20

2.5.1 Rasgos biométricos suaves 21

2.5.2 Datasets soft biometrics 24

3. ENTORNO EXPERIMENTAL 27

3.1 ANÁLISIS DE LA BASE DE DATOS 27

3.2 RASGOS BIOMÉTRICOS SUAVES SELECCIONADOS 29

3.3 REDES NEURONALES CONVOLUCIONALES PREVIAMENTE ENTRENADAS 31

4. DISEÑO Y DESARROLLO DEL SISTEMA 35

4.1 PREPROCESADO DE LA BASE DE DATOS 35

4.1.1 SEGMENTADO DE LAS IMÁGENES 36
4.1.2 REAJUSTE DE TAMAÑO 37
4.2 TRANSFER LEARNING ALEXNET. 39
4.2.1 Capas de la red Alexnet *39*
4.2.2 Opciones de entrenamiento de la red Alexnet *40*
4.3 TRANSFER LEARNING GOOGLENET 41
4.3.1 Capas de la red GoogLeNet *41*
4.3.2 Opciones de entrenamiento para GoogLeNet *42*
4.4 CONFIGURACIÓN RED CNN PROPIA 43
4.4.1 Capas de la red propia *43*
4.4.2 Opciones de entrenamiento para la red propia *45*
5. INTEGRACIÓN, PRUEBAS Y RESULTADOS 47
5.1 DIVISIÓN DATASET PETA 47
5.2 ENTORNO DE PRUEBAS 50
5.3 RESULTADOS Y ANÁLISIS 59
6. CONCLUSIONES Y TRABAJO FUTURO 67
6.1 CONCLUSIONES 67
6.2 TRABAJO FUTURO 68
REFERENCIAS 69

ÍNDICE DE FIGURAS

FIGURA 2.1: INTELIGENCIA ARTIFICIAL..... 5

FIGURA 2.2: APRENDIZAJE DE MÁQUINA VS APRENDIZAJE PROFUNDO 6

FIGURA 2.3: RED CON UNA CAPA PERCEPTRÓN Y RED NEURONAL MULTICAPA 10

FIGURA 2.4: NEURONA ARTIFICIAL 10

FIGURA 2.5: ALGORITMO BACKPROPAGATION 13

FIGURA 2.6: ESQUEMA DE UNA RED NEURONAL CONVOLUCIONAL 16

FIGURA 2.7: EN ORDEN, IMÁGENES DE LAS BASES DE DATOS ORL, YALE Y PIE. 25

FIGURA 3.1: BASE DE DATOS PETA [30] 28

FIGURA 3.2: RED ALEXNET 32

FIGURA 3.3: MÓDULO INCEPTION DE LA RED GOOGLENET 33

FIGURA 4.1: IMÁGENES DE LA MISMA PERSONA EN PETA DESDE DIFERENTES ÁNGULOS CON CALIDADES Y TAMAÑOS VARIADOS 295X101, 270X100, 288X90X3 Y 264X120X3 36

FIGURA 4.2: SEGMENTADO DE IMÁGENES 37

FIGURA 4.3: REAJUSTE DE TAMAÑO PARA RED ALEXNET 227X227 PIXELES 38

FIGURA 5.1: PORCENTAJES DIVISIÓN DATASET 48

FIGURA 5.2: LEYENDA DE COLORES PARA EL ENTRENAMIENTO DE LA RED 50

FIGURA 5.3: ENTRENAMIENTO RED ALEXNET ATRIBUTO COLOR DE PELO, 30 ÉPOCAS. 51

FIGURA 5.4: ENTRENAMIENTO RED ALEXNET ATRIBUTO COLOR DE PELO, 10 ÉPOCAS. 52

FIGURA 5.5: ENTRENAMIENTO RED ALEXNET ATRIBUTO COLOR DE PELO BALANCEADO, 30 ÉPOCAS. 53

FIGURA 5.6: ENTRENAMIENTO RED GOOGLENET ATRIBUTO COLOR DE PRENDA INFERIOR, 30 ÉPOCAS. .. 54

FIGURA 5.7: ENTRENAMIENTO RED GOOGLENET ATRIBUTO COLOR DE PRENDA INFERIOR, 10 ÉPOCAS. .. 54

FIGURA 5.8: ENTRENAMIENTO RED GOOGLENET ATRIBUTO COLOR DE PRENDA INFERIOR BALANCEADO, 30 ÉPOCAS. 55

FIGURA 5.9: ENTRENAMIENTO RED PROPIA ATRIBUTO COLOR DE PRENDA SUPERIOR SECCIÓN 1, 30 ÉPOCAS. 56

FIGURA 5.10: ENTRENAMIENTO RED PROPIA ATRIBUTO COLOR DE PRENDA SUPERIOR CON AUMENTO DE DATOS, SECCIÓN 1, 30 ÉPOCAS. 58

FIGURA 5.11: ENTRENAMIENTO RED PROPIA ATRIBUTO COLOR DE PRENDA SUPERIOR BALANCEADO CON AUMENTO DE DATOS, SECCIÓN 1, 30 ÉPOCAS. 59

FIGURA 5.12: GRÁFICO RESULTADOS FINALES ATRIBUTOS BALANCEADOS 63

FIGURA 5.13: PRUEBA DE EJEMPLO SOBRE IMAGEN SIMÉTRICA EN EL EJE HORIZONTAL 64

ÍNDICE DE TABLAS

TABLA 2.1 RENDIMIENTO OBTENIDO CON LA BASE DE DATOS PETA [30] 24

TABLA 3.1: CARACTERÍSTICAS BASE DE DATOS PETA 28

TABLA 3.2: ATRIBUTOS BASE DE DATOS PETA 29

TABLA 3.3: ATRIBUTOS SELECCIONADOS EN SECCIÓN 1 30

TABLA 3.4: ATRIBUTOS SELECCIONADOS EN SECCIÓN 2 31

TABLA 3.5: ATRIBUTOS SELECCIONADOS EN SECCIÓN 3 31

TABLA 3.6: ARQUITECTURA INCEPTION DE LA RED GOOGLENET 34

TABLA 4.1: APRENDIZAJE POR TRANSFERENCIA EN LA RED ALEXNET 40

TABLA 4.2: APRENDIZAJE POR TRANSFERENCIA EN LA RED GOOGLENET 42

TABLA 4.3: CAPAS DE LA CNN PROPIA 45

TABLA 5.1: DIVISIÓN DE DATASETS SECCIÓN 1 49

TABLA 5.2: DIVISIÓN DE DATASETS SECCIÓN 2 49

TABLA 5.3: DIVISIÓN DE DATASETS SECCIÓN 3 49

TABLA 5.4: DATASETS BALANCEADOS ATRIBUTO COLOR DE PELO 52

TABLA 5.5: DATASETS BALANCEADOS ATRIBUTO COLOR PRENDA INFERIOR 55

TABLA 5.6: DATASETS BALANCEADOS ATRIBUTO COLOR PRENDA SUPERIOR 58

TABLA 5.7: RENDIMIENTO RED ALEXNET POR ATRIBUTO 60

TABLA 5.8: RENDIMIENTO RED GOOGLENET POR ATRIBUTO 61

TABLA 5.9: RENDIMIENTO RED PROPIA POR ATRIBUTO 62

TABLA 5.10: RESULTADOS PRUEBA DE EJEMPLO CON IMAGEN ROTADA SIMÉTRICAMENTE EN EJE HORIZONTAL 65

1.Introducción

En este primer capítulo se presenta el contenido de este proyecto y se indica la motivación que ha llevado a realizar este Trabajo Fin de Máster. Además, se exponen los objetivos del trabajo y la organización es esta memoria.

1.1 Motivación

La biometría permite que una persona sea identificada y autenticada en base a un conjunto de datos reconocibles y verificables, que son únicos y específicos para cada persona.

Un sistema biométrico es el proceso que permite el reconocimiento de personas mediante unos patrones que determinan la autenticación mediante el uso de sus diferentes características biológicas, como pueden ser la huella dactilar, la retina, el iris, o la mano. Pero también existen características conductuales como el reconocimiento de voz, la autenticidad de una firma o incluso la pulsación sobre teclas de ordenador.

El comportamiento humano es una gran fuente de información biométrica que se puede utilizar para establecer o verificar una identidad precisa. La identificación o autenticación de personas es un requisito básico para prevenir los efectos adversos de las crecientes amenazas de seguridad tanto en el mundo real como en el ciber mundo.

Aunque las contraseñas siguen siendo el mecanismo más común de autenticación de personas, los informes sobre la seguridad de los sistemas tradicionales basados en contraseñas señalan lo fácil que es hoy en día averiguar y, por tanto, romper la seguridad de esas contraseñas.

Para superar estos inconvenientes, los sistemas de autenticación basados en biometría tienen una demanda creciente de muchas aplicaciones de seguridad, como la gestión de control de fronteras, tarjetas de identidad nacionales, pasaportes electrónicos o algo más común, el acceso al teléfono móvil.

Este trabajo, se centrará en los rasgos biométricos suaves dentro de los rasgos físicos y conductuales, como son el género, forma y color de pelo, o prendas de vestir. Estos rasgos nos son exclusivos de un sujeto especifico, pero son útiles para la identificación, verificación y descripción de personas.

Los atributos biométricos suaves se pueden combinar con rasgos biométricos clásicos para mejorar la precisión de un sistema de reconocimiento biométrico o se pueden usar como filtros para restringir la búsqueda durante una operación de identificación biométrica.

El aprendizaje dentro de inteligencia artificial y el aprendizaje profundo con el uso de los atributos biométricos suaves para diferentes aplicaciones de reconocimiento de personas han sido la principal motivación para la realización de este trabajo Fin de Máster.

1.2 Objetivos

El principal objetivo de este Trabajo Fin de Máster es aprender en profundidad las redes neuronales convolucionales y desarrollar un sistema basado en la clasificación de atributos suaves de personas para después compararlos con redes previamente entrenadas.

En primer lugar, se realizará una tarea investigadora. Se estudiará los distintos modelos que existen en el aprendizaje profundo y, más en detalle, las redes neuronales convolucionales. Todo ello, con especial atención en la detección de atributos físicos suaves de personas, que es lo que se analizará en nuestros sistemas.

Tras obtener los conocimientos necesarios se realizarán modificaciones en redes pre entrenadas existentes y comúnmente conocidas. Finalmente se creará una red propia desde cero, con la que poder comparar los resultados para poder realizar una conclusión de todo lo aprendido.

Para todo ello, se utilizará la base de datos *Pedestrian Attribute* (PETA) compuesto por diecinueve mil imágenes de personas de cuerpo entero.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

- Capítulo 1: Introducción. Se presenta la motivación que ha llevado a realizar este proyecto, así como los objetivos marcados al inicio y la organización que tendrá la memoria.

- Capítulo 2: Estado del arte. Se detalla el aprendizaje profundo y las redes neuronales convolucionales, centrándose en las características que cada capa puede aportar para el reconocimiento de atributos corporales a distancia, los cuales serán analizados en el contexto de los rasgos biométricos y trabajos relacionados hasta el momento con la misma base de datos.

- Capítulo 3: Entorno experimental. Se explicará la base de datos que se utiliza en este proyecto, así como las redes neuronales convolucionales previamente entrenadas a las que se les realizará posteriormente un aprendizaje por transferencia.

- Capítulo 4: Diseño y desarrollo del sistema. Se detallará en profundidad el procesado de la base de datos y se explicará el desarrollo realizado en las redes previamente entrenadas y la red propia creada desde el inicio.

- Capítulo 5: Pruebas y resultados. Aquí se expondrán las pruebas realizadas, los resultados obtenidos y el análisis de estos.

- Capítulo 6: Conclusiones y trabajo futuro. En este último capítulo se comprobará si se han cumplido los objetivos marcados y se plantearán futuras líneas de trabajo que puedan mejorar el sistema clasificador de atributos corporales a distancia.

2.Estado del arte

En este capítulo se desarrollará el Estado del Arte que sirve de base para la correcta comprensión de este Trabajo Fin de Máster. En primer lugar, se detallará el aprendizaje profundo. A continuación, se dará un fundamento teórico de las Redes Neuronales Convoluciones, aplicadas a la clasificación, ya que es la base de los sistemas implementados en este trabajo. Y por último, el estado del arte de los rasgos biométricos suaves, los cuales han sido analizados y clasificados en las CNN de este trabajo.

2.1 Aprendizaje profundo

El aprendizaje de máquina o automático es una disciplina científica dentro de la inteligencia artificial que crea sistemas capaces de predecir automáticamente qué situaciones podrían darse o no. Este tipo de aprendizaje permite mejorar la calidad de vida de las personas, ayudando en sus tareas día a día, creando un entorno más seguro y sencillo.

El aprendizaje profundo, un subcampo dentro del aprendizaje de máquina, ha establecido en los últimos años una nueva y emocionante tendencia en el aprendizaje automático pues aborda este problema exacto e implica la creación de programas informáticos complejos que son capaces de aprender y, por lo tanto, mejorar sus actuaciones reuniendo más datos y experiencias.

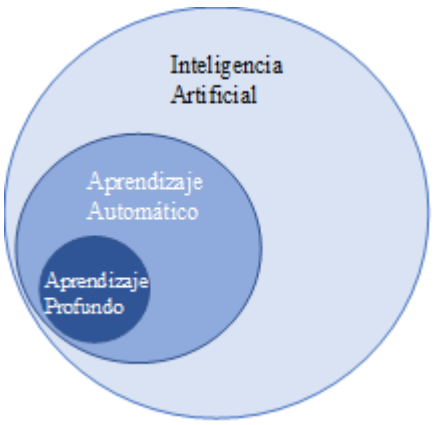

Figura 2.1: Inteligencia Artificial

Mientras que el aprendizaje de máquina utiliza conceptos simples, el aprendizaje profundo utiliza diferentes algoritmos de aprendizaje automático para modelar abstracciones de alto nivel en datos usando arquitecturas jerárquicas, conocidas como redes neuronales profundas.

Estos sistemas tienen que ser capaces por sí mismos de adquirir su propio conocimiento, extraer deducciones, obtener patrones que utilicen esos conceptos simples y que los relacionen entre sí para formar otros más complejos, de forma que exista una jerarquía de conceptos con muchas capas. De hecho, el funcionamiento de estos algoritmos trata de imitar al cerebro humano.

Cada capa procesa la información y arroja un resultado dado en forma de ponderación, asignando un porcentaje o tasa de acierto sobre esa capa. Cada capa utilizará el resultado de la capa previa para modificarla y volver a ponderarla sacando su propia conclusión. Este sistema reduce la tasa de error y aumenta la precisión de las conclusiones.

El esquema principal de un sistema de aprendizaje automático frente a un sistema de aprendizaje profundo se puede observar a continuación en la Figura 2.2, donde dado un dato de entrada, el algoritmo de aprendizaje profundo extrae las características a la vez que clasifica la etiqueta realizando este proceso en cada capa, mientras que a un sistema de aprendizaje automático se le aportan las características encontradas en el dato de entrada y a partir de ahí realiza la clasificación.

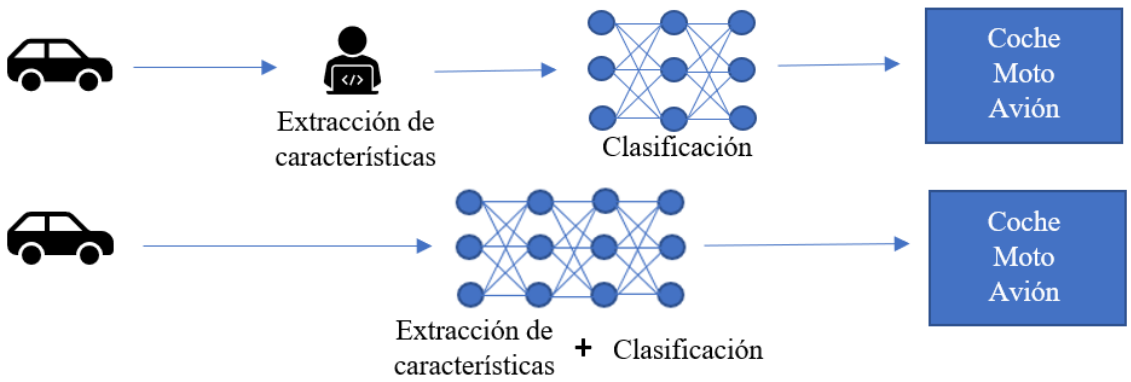

Figura 2.2: Aprendizaje de máquina vs Aprendizaje profundo

Hasta hace unos años, las redes neuronales estaban limitadas por la potencia de los ordenadores, por tanto, estaban limitadas en su complejidad. Además, se necesita una gran cantidad de información para que un sistema de aprendizaje profundo entrene y aprenda. Sin embargo, los avances en el análisis de Big Data han permitido redes neuronales más grandes y sofisticadas, permitiendo a los sistemas de aprendizaje observar, entrenar, aprender y reaccionar ante situaciones complejas más rápido que los seres humanos.

Diferentes estudios han concluido que dado un conjunto de datos pequeño se obtienen mejores resultados en sistemas de aprendizaje automático, pero a medida que se aportan más datos de entrada para entrenar, esos resultados se ven claramente invertidos mejorando con el aprendizaje profundo.

El aprendizaje profundo ha ayudado a la clasificación de imágenes, traducción de diferentes idiomas o reconocimiento de voz. Se pueden utilizar en múltiples campos de investigación para resolver cualquier problema de reconocimiento de patrones sin intervención de personas.

2.1.1 Tipos de aprendizaje profundo

Dentro del aprendizaje profundo, se puede diferenciar tres grandes categorías de aprendizaje [1]. A continuación, se describe cada uno de ellos:

- Aprendizaje supervisado: es el modelo que trabaja con datos de entrada etiquetados previamente, de forma manual o automática. Los algoritmos tratan de encontrar una función que, dadas las variables de entrada, les asigne la etiqueta de salida adecuada. Es decir, este tipo de algoritmos aprende del conjunto de datos proporcionado y finalmente, predice el valor de salida. El aprendizaje supervisado se suele clasificar en:
    - Problemas de clasificación. Su variable objetivo es de tipo categórico. Por ejemplo, un sistema de identificación de dígitos
    - Problemas de regresión. Su variable objetivo es de tipo numérico como, por ejemplo, predicciones meteorológicas.

  Algunos de los algoritmos más habituales dentro de este tipo de aprendizaje son los Árboles de Decisión, regresión logística, clasificación de Bayes, Máquinas de Vectores de Soporte y Redes Neuronales Convoluciones. Estas últimas se definirán más adelante en este capítulo pues es lo utilizado en este Trabajo Fin de Máster.

- Aprendizaje no supervisado: es el modelo que trabaja el entrenamiento con datos de entrada sin etiquetar previamente. Este modelo se ajusta a partir de las observaciones al carecer de un conocimiento a priori. Por tanto, solo se puede describir la estructura de los datos para intentar encontrar algún tipo de organización que simplifique el análisis. Se dice que este modelo tiene un carácter exploratorio. El Aprendizaje no supervisado se suele usar en problemas de clustering, agrupamientos de co-ocurrencias y perfilado o *profiling*.

  En las tareas de clustering, se buscan agrupamientos basados en similitudes, pero nada garantiza que éstas tengan algún significado o utilidad. Explorar datos sin un objetivo definido puede dar de resultado correlaciones espúreas curiosas, pero poco prácticas.

- Aprendizaje por refuerzo: es el modelo donde el algoritmo de aprendizaje recibe algún tipo de valoración acerca de la idoneidad de la respuesta dada, distinguiendo cuando la respuesta es correcta, si se le da la información de que ha sido apropiada, o si la respuesta es errónea que entonces solo le informa que el comportamiento es incorrecto. Esta información que se aporta sirve como *feedback* o

retroalimentación. Por tanto, es un sistema que aprende a base de ensayo-error. Muchos videojuegos se consideran de este modelo, como el Space Inviders.

2.1.2 Conjunto de datos o Data Set

Los algoritmos de aprendizaje profundo son sistemas muy útiles dentro de la inteligencia artificial, donde la intervención del ser humano es prácticamente insignificante y los resultados obtenidos son casi comparables con la inteligencia humana. Pero para que esto sea posible, es necesario aportar al algoritmo gran cantidad de información. A más información introducida en el sistema, mejores resultados se obtienen.

Muchos conjuntos de datos no son de uso público, por eso, a la hora de realizar un aprendizaje profundo es necesario recopilar una base de datos muy extensa, preferiblemente etiquetada previamente, o tener los medios para crear de cero una base de datos de este tipo.

Además, muchos conjuntos de datos públicos se están volviendo obsoletos rápidamente en términos de tamaño y densidad, pues para los modelos existentes en la actualidad han sido necesarios del orden de millones de datos etiquetados para que los sistemas aprendan a partir de funciones complejas y realicen una transformación automática de estos datos.

Aunque ha habido un progreso notable en la mejora de los algoritmos de aprendizaje profundo y en el desarrollo de sistemas de entrenamiento de alto rendimiento, faltan avances en la construcción del conjunto de datos. El tamaño de los conjuntos de datos para capacitación y evaluación no está aumentando mucho, sino que está rezagado y obstaculizando el progreso en el aprendizaje profundo [2].

2.2 Redes Neuronales Artificiales

N. Singh Chauhan definió las redes Neuronales Artificiales (RNA) como “un modelo matemático que está basado en las redes neuronales biológicas y por lo tanto es similar al funcionamiento de un sistema biológico neuronal” [3].

Una red neuronal artificial es un tipo de modelo de aprendizaje automático que consiste en el uso de nodos, también llamados neuronas, caracterizado por su inspiración en las estructuras biológicas a las que hace referencia. Estos nodos conectados entre sí se adaptan al proceso de aprendizaje de la red, proporcionando un valor a la salida de cada nodo dependiendo de la entrada.

En comparación con los algoritmos convencionales, las redes neuronales pueden resolver problemas bastante complejos. Esto, unido a su estructura simple y organizada que permite abordar multitud de problemas, se convierte en la razón principal para usarlas.

2.2.1 Breve historia de las redes neuronales

Las primeras investigaciones en redes neuronales se realizaron hace más de 60 años. Entre las décadas de 1950 y 1960 el científico Frank Rosenblatt inspirado en el trabajo de Warren McCulloch y Walter Pitts creó el Perceptrón, la unidad donde nacería y se potenciarían las redes neuronales artificiales [4].

Rosenblatt dedujo que, para calcular la salida binaria de la neurona, debía introducir el concepto de pesos, un valor que expresa la importancia de la respectiva entrada con la salida. Sus principales usos son decisiones binarias sencillas o funciones lógicas como OR y AND.

En 1965 comenzó a investigarse el perceptrón multicapa, donde se iniciaron las capas ocultas, pero manteniendo las entradas y salidas binarias. Seguían calculando manualmente los pesos, y cuantos más perceptrones en cada capa, más difícil era calcular los pesos para obtener las salidas.

Para que las redes neuronales aprendieran solas fue necesario introducir las neuronas Sigmoideas en 1980, permitiendo que las entradas tuvieran más valores, además de 0 o 1, realizando pequeñas alteraciones en valores de pesos y, por tanto, en la salida. Además, se añadió el termino *bias* o sesgo para que las neuronas de una misma capa sumasen 1. Con estas aportaciones se creó la primera función de activación.

En los años posteriores, se fue creando los términos de *Feedforward* y *Backpropagation*, pero fue en 1989 cuando Yann LeCun creó la primera red neuronal convolucional enfocada en el reconocimiento de letras manuscritas.

La arquitectura constaba de varias capas que implementaban la extracción de características y luego las clasificaban. Esta arquitectura que usaba capas profundas y realizaba una clasificación de la salida ofrecieron un mundo nuevo de posibilidades en las redes neuronales.

Durante muchos años no se pudo alcanzar la profundidad de este aprendizaje por la falta de poder de cómputo. Es a partir de 2006 que se supera esta barrera gracias a las Unidades de Procesamiento Gráfico, GPU, que permiten una capacidad casi ilimitada en complejidad de red.

2.2.2 Estructura de una red neuronal

Cada red neuronal está formada por diferentes capas que a su vez están formadas por nodo/s. Los nodos se encuentran interconectados con los nodos de la capa previa y posterior.

La capa de entrada recibe información de fuentes externas, como los valores de los atributos, llamada datos de entrada a la red. Y la capa de salida o última capa, produce la salida de la red. Entre medias se encuentran las capas ocultas. El número de capas varía en función de la tipología [3].

En la Figura 2.3 se puede observar la diferencia entre dos redes neuronales, una posee una única capa oculta, también llamada Perceptrón, la otra tiene más de una capa oculta, también llamada red multicapa. Esta es una de las formas que se puede clasificar una red neuronal, si posee una o más capas ocultas.

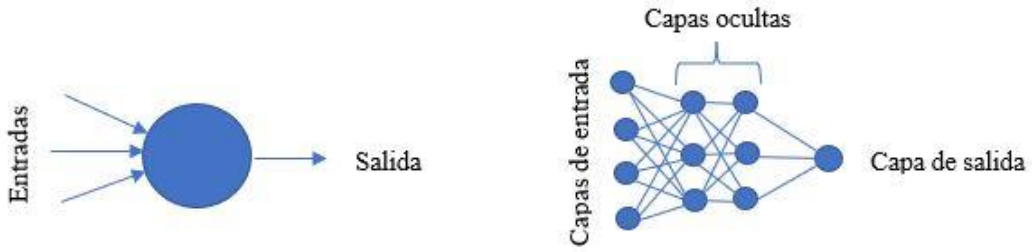

Figura 2.3: Red con una capa perceptrón y Red neuronal multicapa

Como se ha mencionado, una red neuronal tiene tres tipos de capas:

- De entrada: es la capa que recibe directamente la información proveniente de las fuentes externas a la red.
- Ocultas: son internas a la red y no tienen contacto directo con el entorno exterior. Pueden estar interconectadas de distintas maneras, lo que determina junto con su número las distintas tipologías de la red. Si la red tiene más de dos capas ocultas es considerada una red neuronal profunda.
- De salida: transfieren información de la red hacia el exterior.

Además de la complejidad en el algoritmo, cuantas más capas ocultas posea, más aumenta el tiempo de cómputo.

Cada capa está formada por neuronas artificiales o nodos. Cada neurona tiene varias entradas y una salida. En la siguiente Figura, se puede observar mejor como funciona cada neurona en una capa oculta, imitando el comportamiento que tiene una neurona biológica.

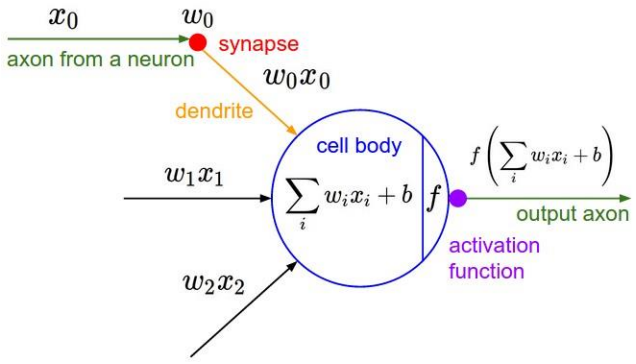

Figura 2.4: Neurona artificial

En ella un sesgo constante se añade al sumatorio del producto escalar de las entradas a la neurona y su matriz de pesos. Cada conexión tiene asociado un peso, que se trata de un número que se va ajustando a lo largo del entrenamiento de dicha red neuronal.

Inicialmente, es necesario indicar los pesos de manera aleatoria para no influir en el proceso de entrenamiento. Pero después de eso, el peso va variando según la información que aporta la capa previa y que consigue alterar considerablemente el resultado final de la red.

Finalmente, para obtener el resultado de la neurona en la capa oculta, se pasará por una función de activación que añadirá la no-linealidad a la red.

A este algoritmo lineal en el cual se basan las redes neuronales y que adquiere una modificación en el resultado tras cada capa oculta, se le conoce como algoritmo *Feedforward* [5].

2.2.3 Funciones de activación

Existen diferentes funciones de activación que determinan si aplican una linealidad o no-linealidad a la red neuronal [3].

La salida de una red neuronal que no usa una función de activación será simplemente una función lineal, o lo que es lo mismo, un polígono de grado uno. Si se utiliza una función de activación lineal, la red solo puede adaptarse a los cambios lineales de entrada. Pero la realidad es que los errores poseen características no lineales que, sumados a las redes neuronales, tienen la capacidad de aprender de estos datos erróneos.

La propiedad más atractiva de las redes neuronales artificiales es la capacidad de adaptar su comportamiento de acuerdo con las características que van cambiando en el sistema.

Esto implica que, aunque una ecuación lineal es más fácil de resolver, su complejidad es limitada y no tiene la capacidad de aprender datos complejos, lo que significa que tendrá un rendimiento y potencia limitados. Por eso, es preferible que una red neuronal aprenda de datos más complejos.

A continuación, se describen las funciones de activación más utilizadas:

- **Función escalón binario**: tiene un umbral clasificador que permite activar la neurona si el valor obtenido en ella es superior a ese umbral, de lo contrario, la neurona se desactiva y no será entrada a la siguiente capa. Es una función lineal y no se puede utilizar para redes multiclases.

$$f(x) = 1, x \geq 0$$
$$f(x) = 0, x < 0 \tag{2.1}$$

- **Función lineal**: función lineal como su nombre indica, es directamente proporcional a la entrada. La constante es elegida por el usuario.

$$f(x) = ax \tag{2.2}$$

- **Función Sigmoide**: también conocida como la función logística, se trata de una función continuamente diferenciable y no lineal. La aplicación de esta funciona da

como resultado un aprendizaje más rápido de la red y proporciona una prevención de efectos de sobrecarga.

El efecto de sobrecarga se produce cuando uno o varios atributos obtienen una gran influencia en el atributo a predecir, haciendo que otros no puedan aportar apenas información.

Esta función da como resultado un rango comprendido entre 0 y 1 al estar normalizados los valores. Se caracteriza por tener una forma curva como una "S".

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.3}$$

- **Función tangente hiperbólica tanh**: también tiene forma de "S" pero las salidas tendrán unos valores comprendidos entre -1 y 1, simétrica en el origen. Además, solo las entradas de valor cero se convierten en salidas cercanas a cero, esto hace prevenir a la red de quedarse sin avanzar durante el entrenamiento como sí puede ocurrir con la función Sigmoide. Esta función no lineal se usa principalmente para la clasificación entre dos clases, al igual que pasa con la función sigmoide.

$$f(x) = \frac{e^{x} - e^{x}}{e^{x} + e^{x}} \tag{2.4}$$

- **Función ReLU**: su nombre proviene de las siglas de *Rectified Linear Unit*, aunque también es conocida como la función rampa. Permite el paso de todos los valores positivos sin alterarlos, pero asigna todos los valores negativos a cero. Es la más usada entre las no lineales actualmente porque es usada en prácticamente todas las redes neuronales convolucionales o aprendizaje profundo.

$$f(x) = \max(0, x) \tag{2.5}$$

- **Función Softmax**: es la combinación de múltiples funciones sigmoides, lo que permite su uso en problemas multiclase. Es no lineal y su salida, también se encuentra comprendida entre 0 y 1.

$$f(x_i) = \frac{e^{x_i}}{\sum_i e^{x_i}} \tag{2.6}$$

En el pasado, las funciones de activación más frecuentes eran Sigmoide y Tanh, pero actualmente se conoce que tiene mejor resultados la función ReLU porque evita el problema llamado fuga de gradiente.

Este problema se describe como la situación donde una red multicapa *Feedforward* no puede propagar información útil del gradiente y produce que el modelo converja prematuramente a una solución deficiente.

Una característica importante de una función de activación es que debe ser diferenciable para que se pueda implementar el algoritmo *Backpropagation* y así, reducir errores.

2.2.4 Algoritmo Backpropagation

Otro problema que surge en las redes neuronales es ajustar los pesos, para que la función de coste sea minimizada. Para calcular los gradientes, se utiliza el algoritmo de *Backpropagation* o retro propagación.

El algoritmo de *Backprogation* [3] reajusta los pesos de las interconexiones en la red neuronal, basado en tasas de errores locales. Esto significa que después de que se haya hecho una predicción para un conjunto de valores de entrada, el valor real de salida se compara con el valor de predicción y se calcula el error. Este error es usado para reajustar los pesos de las conexiones comenzando por los extremos que están conectados directamente a los nodos de salida de la red, para ir avanzando hacia atrás. Esto consigue que la red tenga el rendimiento deseado y, por tanto, mejores resultados.

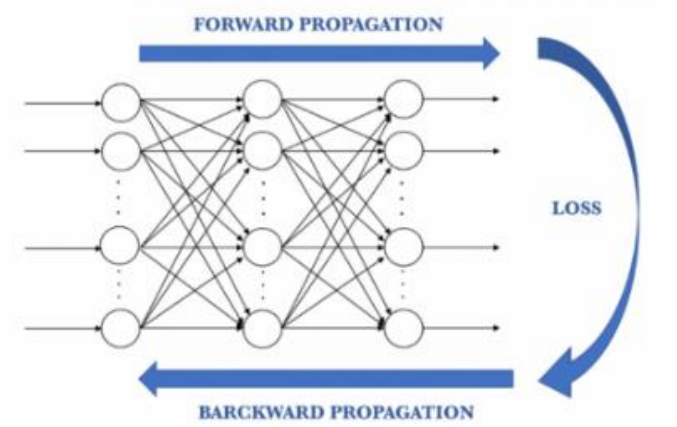

Figura 2.5: Algoritmo backpropagation

2.2.5 Función de Coste

Basada en la diferencia entre el valor actual y el valor predicho, el valor de error llamado Función de Coste es calculado y devuelto al sistema para obtener unos resultados óptimos.

Para cada capa de la red, la función de coste es analizada y usada para ajustar el umbral y los pesos de la siguiente entrada. El objetivo es minimizar la función de coste. Cuanto más bajo sea este valor, más cerca se encuentra el valor actual del valor a predecir. De esta forma, el error sigue disminuyendo marginalmente en cada ejecución a medida que la red aprende a analizar los valores. Se alimentan los datos resultantes a través de toda la red neuronal.

Una vez ajustado este valor y ejecutada la red de nuevo, la diferencia entre ambos valores debe ser más pequeña, es decir, la función de coste debe tener un valor inferior al previo. Este proceso se debe repetir hasta reducir el Coste al máximo posible.

Existen dos métodos principales para ajustar los pesos [3]:

- **Brute-Force ;ethod**: es el más utilizado para redes neuronales con una capa oculta donde se toman un conjunto de posibles pesos iniciales y a medida que se va ejecutando el algoritmo, se eliminan los menos óptimos, quedando finalmente con el mejor peso para el coste mínimo.
- **Batch-Gradient Descent**: es el método apropiado para redes neuronales multicapa. Es un algoritmo de optimización iterativa que observa el recorrido de la pendiente, o gradiente, obtenida en un punto. De tal manera que, el conjunto de datos obtenidos en el entrenamiento se tiene en consideración para cada paso. Se realiza la media de todos los gradientes del entrenamiento completo y con ese valor se actualizan los parámetros de esa neurona.

2.2.6 Stochastic Gradient Descent (SGD)

Para poder abordar un problema de cálculo de coste cuando se tiene gran cantidad de datos, el proceso de *Batch Gradient Descent* no es eficiente pues el cálculo de todos los gradientes puede tomar demasiado tiempo.

Para ello, comúnmente se utiliza el algoritmo *Stochastic Gradient Descent* [3] que actualiza los pesos de la red de forma iterativa en función de los datos de entrenamiento, pero fijando previamente el parámetro tasa de aprendizaje o *learning rate*.

La tasa de aprendizaje se encuentra entre 0 y 1, y es multiplicado por el error local de cada valor de salida. Para el entrenamiento de la red se fija este valor, aquí se realiza una tarea de prueba-error donde tras ajustar el valor se decide si minimizarlo, normalmente de forma exponencial, para mejorar el resultado, pero teniendo en consideración el tiempo. Pues a menor tasa de aprendizaje, mejor resultado de convergencia, pero peor tiempo de entrenamiento de la red.

2.3 Redes neuronales convolucionales

Dentro de los algoritmos de aprendizaje profundo, en particular uno ha ido avanzando y perfeccionándose en mayor medida pues ha contribuido en el desarrollo y perfeccionamiento del campo de Visión Artificial, las redes neuronales convolucionales, CNN.

Gracias a ellas, desde apenas 1998, podemos clasificar imágenes, detectar diversos tipos de tumores automáticamente, enseñar a conducir a coches autónomos o identificar personas entre un sinfín de aplicaciones.

Las CNNs son una poderosa técnica de aprendizaje automático del campo del aprendizaje profundo. Estas redes se entrenan utilizando grandes colecciones de imágenes, de las que pueden aprender ricas representaciones de características.

Las redes neuronales convolucionales utilizan imágenes como información de entrada a la red, tomando estas como input y asignando pesos a ciertos elementos en la imagen para así poder diferenciar unos de otros.

Las redes convolucionales son multicapa, donde cada capa detecta diferentes aspectos de la imagen y extrae características con distintos niveles de abstracción, hasta poder reconocer formas complejas como rostros o siluetas.

La capa de salida realizará la clasificación de la imagen entre las etiquetas aportadas a la entrada de la red.

Este tipo de red neuronal debe aprender por sí sola a reconocer una diversidad de objetos dentro de imágenes y para ello se necesita gran cantidad de información para entrenar.

La red toma como entrada los pixeles de una imagen, siendo el total de pixeles el número de neuronas de entrada. Una imagen en escala de grises multiplica por la unidad el número de pixeles, mientras que una imagen a color necesita tres canales RGB y por tanto multiplica el número de neuronas necesarias por tres.

2.3.1 Preprocesamiento y feature mapping

Para facilitar el computo de la red, los valores de los pixeles se normalizan, pasan de tener un valor comprendido entre 0 y 255 a un valor ente 0 y 1.

El proceso consiste en aplicar la convolución entre los pixeles de la imagen de entrada y un kernel, o tres kernel para imágenes a color. Al conjunto de kernels que son utilizados se le llama filtro. Esto permite identificar bordes o enfoques, entre otros aspectos.

El filtro recorre el total de las neuronas de entrada a la red para formar, finalmente, una de las capas ocultas. Este procedimiento se conoce con el nombre de *feature mapping*.

Este tipo de redes se forma con tres tipos de capas específicas:

- **Capas convolucionales**: Reciben de la capa de entrada la imagen, a la que aplican los filtros realizando la convolución entre todos los pixeles y devolverá una nueva imagen que ha extraído un mapa de características o patrón de la imagen original.
- **Capas de reducción o *pooling***: Colocada tras la capa convolucional. Se encarga de reducir las dimensiones espaciales para la entrada a la siguiente capa de convolución manteniendo únicamente donde prevalecen las características más importantes que detectó cada filtro. La operación más común es *max-pooling*, donde la salida es el máximo de la entrada en ventana, reduciendo así el número de parámetros.
- **Capas totalmente conectadas o *fully conected***: Es la capa final, donde todos los nodos de la capa están interconectados. Es la capa clasificadora, que tendrá el mismo número de neuronas que clases a predecir. Existen dos tipos de situaciones donde puede encajar una capa *fully-connected*, si se necesita aplanar la salida de la capa previa para convertirla en un vector único para la entrada de la siguiente capa, o si se espera que dé el resultado final con las probabilidades para cada clase.

Las imágenes tienen una altura, una longitud y una profundidad. Esta última característica será tres si la imagen es en color, o uno si es en blanco y negro. Tras pasar la capa de entrada por una capa convolucional y de reducción se verá disminuida en tamaño

de altura y longitud, pero aumentará la profundidad, pues por cada filtro que crea se extraen diferentes características de la imagen.

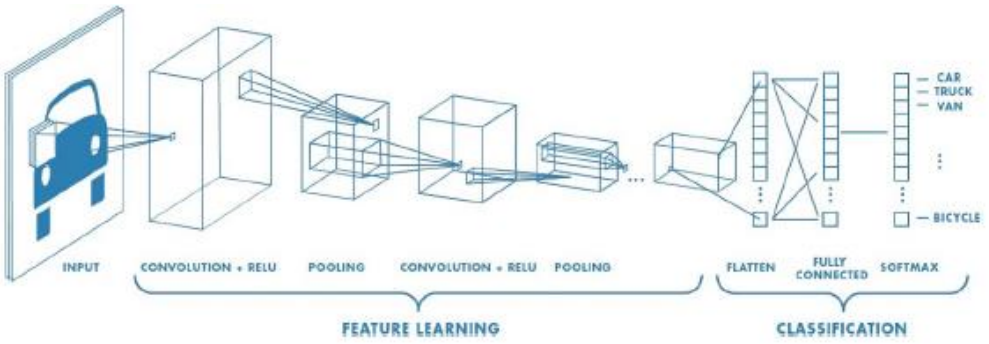

Figura 2.6: Esquema de una red neuronal convolucional

Antes de realizar la convolución, es recomendable valorar la base de datos que se posee por si algunas mejoras se pueden implementar para facilitar a la red la extracción de características. Existen los siguientes parámetros que se deben tener en cuenta en la capa convolucional antes de comenzar el entrenamiento de una red.

- *Stride*: parámetro que indica cada cuanto se quiere recorrer el filtro a lo largo de la imagen. Por defecto *Stride*=1. Cuanto mayor sea este parámetro, más se reduce el tamaño de la imagen.
- *Padding*: esta operación consiste en agregar pixeles con valor cero alrededor de la imagen original. Se utiliza por dos motivos: para que la imagen resultante sea de igual tamaño que la imagen original; y para no perder información relevante de las esquinas de la imagen original, pues al realizar la convolución, los valores de los pixeles centrales toman mayor relevancia.

2.3.2 Aprendizaje por transferencia

Entrenar una red neuronal convolucional para que sea capaz de clasificar una imagen se puede realizar creando una red desde cero, o realizar el proceso de aprendizaje por transferencia sobre una red existente.

De la misma forma que los seres humanos somos capaces de aplicar lo aprendido en una tarea para aplicarlo a otra similar, el proceso conocido como aprendizaje por transferencia o *fine-tuning* consiste en la inicialización de una red previamente entrenada para clasificar un tipo de imágenes, para después volver a entrenarla sobre un conjunto nuevo de datos con etiquetas nuevas y que sea capaz de realizar una nueva clasificación [6] [7].

La red se aprovecha tomando el conocimiento adquirido previamente, además de sus características y pesos. Este ajuste permite mejorar significativamente la capacidad de adaptación dado un conjunto de datos etiquetados previamente.

El éxito que proporciona utilizar redes pre entrenadas con *fine-tuning* se produce principalmente al haber sido inicialmente entrenadas con un conjunto de datos mayor del que se dispone.

2.3.3 Entrenamiento de la red

El proceso de entrenamiento de la red consiste en utilizar un conjunto de imágenes etiquetadas para que una red con diferentes capas ocultas aprenda. Es un proceso lento y pueden influir algunos factores, además de las capas de la red, las opciones de entrenamiento.

Se debe tener un conjunto de datos lo más grande posible, pero esto no siempre es posible, por eso existen algunas técnicas que modifican ligeramente las imágenes que se posee para que la red tome estas variaciones como imágenes diferentes. A esta técnica se le conoce como aumento de datos.

El número de neuronas de la capa de entrada corresponderá con el número de clases que se quiere clasificar de un atributo.

Para determinar el número adecuado de capas ocultas, se debe valorar el rendimiento con diferentes arquitecturas. Se ha demostrado que no por tener más capas ocultas, los resultados son mejores, lo necesario es ajustar los parámetros de estas.

Una de las principales opciones de entrenamiento es la tasa de aprendizaje, es necesario ajustar este valor para que la función de error logre la convergencia pero que la red no tenga una velocidad de convergencia muy lenta.

El ultimo parámetro importante es el número de épocas que se debe ajustar para cada red pues no existe un número ideal de épocas en el entrenamiento. No se debe exceder configurando este parámetro con un número muy grande porque puede sobreajustar el modelo.

2.4 Reconocimiento biométrico

Según la RAE, se define biometría como “el estudio mensurativo o estadístico de los fenómenos o procesos biológicos”. La biometría aplicada a la identificación de personas es el método automático de reconocimiento de una persona basado en características fisiológicas o de comportamiento.

Los sistemas creados para la identificación de personas dada una imagen pueden conseguir unos resultados muy satisfactorios para escenarios controlados. Pero todo cambia cuando el escenario varía porque multitud de variables pueden influir en la correcta identificación.

Cualquier persona lleva a cabo una identificación biométrica en su día a día, utilizando esta técnica de manera inconsciente para reconocer a una persona. Los diferentes sistemas

de reconocimiento biométricos que existen actualmente varían en complejidad o modo de funcionamiento, pero la finalidad es la misma, verificar o establecer la identidad de una persona.

En ambos casos, tanto personas como sistemas de reconocimiento hacen uso de una información previamente adquirida y son capaces de adaptar esa información a cada caso particular.

Dado que los principales sistemas de identificación de personas han sido creados con el fin de aumentar la seguridad, el proceso evolutivo de la biometría sigue avanzando día a día para que la sociedad disponga de mejores sistemas de seguridad.

En la época en la que vivimos esto es necesario, porque vivimos en un mundo donde existen más dispositivos móviles con acceso a internet que personas en la Tierra. Encontrar un sistema infalible e inequívoco para reconocer personas es el principal objetivo de la biometría

2.4.1 Ventajas de la biometría

Las principales ventajas que presenta la biometría es que es más segura y cómoda frente a sistemas tradicionales como pueden ser contraseñas, patrones, tarjetas, llaves u otros que pueden ser susceptibles de ser transferidos, sustraídos, descifrados o falsificados con fines fraudulentos.

Sin duda, la biometría puede ser una alternativa o complemento en técnicas de identificación y autenticación ya existentes. A continuación, se describen los múltiples beneficios que presenta frente a técnicas tradicionales de identificación. [8]

- Comodidad del usuario: No es necesario recordar diferentes contraseñas o llevar consigo una tarjeta o llave ya que se utilizan elementos de identificación intrínsecos en las personas, no elementos externos.
- Necesidad de secreto: Las contraseñas deben ocultarse y las tarjetas no pueden estar al alcance de otras personas, mientras que la biometría no requiere de tantas medidas de seguridad debido a que resulta más complejo de falsificar. Además, las contraseñas suelen tener relación con la vida del individuo como fechas importantes o nombres de familia, y esto implica mayor vulnerabilidad.
- Posibilidad de robo o perdida: Un rasgo biométrico es mucho más complejo de robar que una llave o contraseña. Además, los rasgos biométricos permanecen invariables salvo situaciones excepcionales.
- Proceso de comparación: Comparar dos rasgos biométricos requiere de mayor capacidad computacional que comparar dos contraseñas.
- Vulnerabilidad frente a un ataque por fuerza bruta: Mientras que una contraseña tiene un número de caracteres razonable, un rasgo biométrico implica cientos de bytes, consiguiendo ser más efectivo frente a ataques por fuerza bruta.

- Autenticación del usuario: Un rasgo biométrico es intransferible, no puede ser prestado o compartido como una llave.
- Coste: A pesar de ser cierto que la implantación de un sistema de reconocimiento biométrico es más costosa que un sistema de contraseñas o una tarjeta física, el coste de mantenimiento es menor ya que no conlleva gastos asociados a pérdidas, olvidos o robos.

Aunque existan múltiples ventajas por las que usar sistemas de reconocimiento biométrico en seguridad, siempre será mejor combinar las tecnologías tradicionales con sistemas profundos de aprendizaje para reforzar los posibles ataques o robos de identidad [9].

2.4.2 Partes de un sistema biométrico

Cualquier elemento significativo de una persona, ya sea un rasgo físico, biológico o conductual, se puede emplear para la identificación de un individuo.

Un equipo biométrico es aquel que tiene capacidades para medir, codificar, comparar, almacenar, transmitir y/o reconocer alguna característica de una persona, con un determinado grado de precisión o fiabilidad.

El funcionamiento de estos sistemas implica la necesidad de un potente software que sea capaz de determinar un grado de precisión alto en la identificación. Algunos de los campos que forman parte de estos sistemas son reconocimiento de formas, inteligencia artificial o complejos algoritmos matemáticos.

Sin embargo, a la hora de desarrollar un sistema de identificación biométrica dado uno o varios elementos de individuos, debe seguir cinco fases diferenciadas:

- **Recolección de datos**: Los sistemas biométricos comienzan con la medida de una característica distintiva de comportamiento o fisiológica de una persona. Tras recopilar la información, se almacenará y formarán el conjunto de datos que caracteriza a ese usuario.

  En esta fase, se incluye la transmisión de datos puesto que muchos sistemas biométricos recogen la información, pero requiere de la transmisión de esta a otra ubicación donde se almacenan y procesan. Aquí se tiene en cuenta el filtrado o compresión de información útil que necesita el sistema, para requerir poco ancho de banda y almacenamiento. Es recomendable no abusar de mecanismos de compresión/descompresión porque generalmente producen pérdida de calidad y los datos pierden información.

- **Procesado de señal:** Cuando ya se tiene la información, se realiza un preprocesado de los datos adecuándolos para facilitar el tratamiento de estos. Algunas de sus funciones pueden ser eliminar el ruido de fondo, normalizar los datos o extraer los bordes de una imagen.

- **Extracción de características**: Los datos son procesados y un conjunto de características son extraídas para representar los rasgos medidos. Esta información es almacenada en una base de datos para su posterior uso.
- **Comparación**: Una vez extraídas las características de la muestra se comparan con las previamente almacenadas dando lugar a una puntuación o probabilidad de semejanza.
- **Decisión**: Tras obtener un resultado de la comparación se valorará el éxito o el fracaso de esa comparación dado un umbral de tolerancia.

En la fase de decisión, se pueden distinguir dos tipos de errores según el umbral fijado:

- Falsa Aceptación (FA): si un individuo se hace pasar por otro y le sistema lo detecta como auténtico, genuino.
- Falso Rechazo (FR): se produce cuando el sistema rechaza a un individuo auténtico como si fuese un impostor.

Este umbral será fijado según el grado de seguridad que se le quiera dar al sistema y evaluando entre poder tener una tasa de falsa aceptación o una tasa de falso rechazo.

2.4.3 Modos de operación

Hasta ahora se ha estado hablando siempre de Identificación Biométrica, sin embargo, la identificación se puede realizar basándose en dos modos de funcionamiento del sistema biométrico:

- **Reconocimiento**: se identifica a un usuario dentro de todos los usuarios que se encuentran almacenados en la base de datos del sistema. El resultado de la comparación puede ser positivo al tener la tasa de probabilidad más alta, o negativo si no supera un determinado umbral.
- **Autenticación**: También llamado verificación. Se compara los datos extraídos de ese usuario con los datos almacenados de la persona que dice ser.

2.5 Rasgos biométricos

Dada toda la información que se puede extraer de una persona entre rasgos biofísicos y conductuales, existen multitud de rasgos biométricos que permiten identificar a una persona.

Los rasgos biométricos más utilizados se pueden clasificar en dos tipos:

A) Biometría estática: Incluye todas las características corporales de las personas. Se pueden dividir en cinco grandes grupos.
   - Reconocimiento de la huella dactilar [10]
   - Reconocimiento facial [11] [12] [13]
   - Reconocimiento de iris/retina

- Geometría de dedos/mano
- Reconocimiento de la firma [14] [15]

B) Biometría dinámica: Incluye las características de comportamiento de las personas. Las más utilizadas son las siguientes:
   - Dinámica del tecleo [16]
   - Reconocimiento de voz [17]
   - Forma de caminar [18] [19] [20]
   - Reconocimiento de escritura [21]

Todas estas características permiten reconocer o autenticar a una persona con una tasa de acierto muy elevada debido a la exclusividad de los patrones que se pueden extraer de ellas. Pero conseguir extraer o almacenar este tipo de características no resulta fácil, pues en la mayoría de los casos se debe tener al individuo muy cerca de sensores, escáneres o micrófonos.

2.5.1 Rasgos biométricos suaves

En 1883, el oficial de policía francés Alphonse Bertillon propuso el método antropométrico, compuesto por un conjunto de medidas del cuerpo humano y marcas individuales con el objetivo de identificar a los delincuentes de forma precisa. Todos los atributos que podía describir a un individuo valían para la identificación de personas dada la falta de automatización en esa época.

Actualmente, a ese conjunto de atributos se les denominan rasgos biométricos suaves. Estos son los rasgos físicos y de comportamiento que pueden ser descritos semánticamente por humanos.

Estos atributos normalmente se obtienen de datos biométricos primarios, son clasificables en categorías predefinidas, comprensibles para el ser humano y se pueden extraer de forma automática. Aunque no posean suficiente distintividad o unicidad para permitir un reconocimiento con alta precisión, se pueden utilizar para mejorar el rendimiento de sistemas biométricos. Además, se pueden considerar un conjunto de ellas formando una bolsa de biometría suave, aumentando la tasa de precisión para identificar al individuo.

Existen diferentes categorías de rasgos biométricos suaves según su naturaleza:

- Demográfico: Edad, genero, etnia, ojos, pelo, color de piel, marcas de la cara.
- Antropométrico y geométrico: Geometría de la cara y el cuerpo.
- Condiciones médicas: Índice de masa corporal, peso o arrugas entre otros.
- Material y comportamiento: Como pueden ser sombreros, gafas, tatuajes, ropa o calzado.

Dado que las características de una persona deben ser los menos cambiantes a lo largo del tiempo para un mejor reconocimiento, algunos atributos suaves son más adecuados,

como pueden ser el género o la etnia [22], que otros que son mucho más variables como pueden ser las prendas de ropa.

El conjunto de datos biométricos primarios se ha estudiado por décadas debido a su amplia gama de aplicaciones. A pesar de haber obtenido buenos resultados en entornos controlados, cuando se dispone de imágenes con baja resolución y a distancia, los resultados son insatisfactorios.

La capacidad de reconocer atributos de viandantes, como puede ser el género o el estilo de ropa, en la lejanía es de interés práctico principalmente en escenarios de videovigilancia donde primeros planos de cuerpo o cara apenas están disponibles.

El reconocimiento de atributos debe realizarse a gran distancia usando la apariencia de cuerpo completo, donde puede ser parcialmente ocluido e incluso en ausencia de rostro, que es la parte del cuerpo donde más información se puede extraer [23].

Existen dos desafíos fundamentales al tratar con rasgos biométricos a gran distancia:

1) Diversidad de apariencia: Debido a las diferentes ropas que llevan los peatones o las variaciones incontrolables de múltiples factores como la iluminación, cámara o el ángulo de visión, puede suceder que un mismo atributo sea clasificado de forma diferente.

   Por ejemplo, una imagen frontal de un viandante puede clasificarse que no lleva mochila, sin embargo, si la imagen es tomada desde el perfil o la parte posterior, se podrá apreciar que lleva una mochila a la espalda.

2) Ambigüedad de apariencia: Dado que los atributos tienen que ser reconocidos en la lejanía, presenta una dificultad añadida debida a la ambigüedad visual inherente y la mala calidad de las características extraídas.

Por estos motivos, aprender a detectar estos atributos de forma automática requiere de una gran base de datos y que estén tomadas de diferentes fuentes, pues confiar en una única fuente y datos de entrenamiento a pequeña escala conduciría fácilmente a un modelo poco realista.

En trabajos previos realizados en este campo de investigación, Hu *et al* [24], presentaron una extracción de características con CNN utilizando diferentes bases de datos, y mostraron que ningún sistema de clasificación de rasgos suaves puede competir con una clasificación de atributos hecha a mano por personas. Paisitkriangkrai *et al* [25], confirmó que las CNN obtenían una clasificación de rasgos biométricos suaves peor que si fuesen clasificados por una persona, y que las redes previamente entrenadas con la base de datos ImageNet, utilizada en el desafío ILSVRC, no consideraban el color como atributo importante. Más tarde, Wu *et al* [26] concluyeron que con una extracción de características manual complementando a las CNN, se obtenían unos mejores resultados.

Recientemente se ha explorado la extracción de atributos de peatones únicamente por medio de CNN [27] [28] [29] . Zhu *et al* [28] aplicó las CNN a la verificación de personas, demostrando que las primeras capas son más sensibles a la semántica de las imágenes,

mientras que las capas intermedias son las encargadas de extraer patrones de bajo nivel como colores o gradientes.

Deng *et al* [30], los creadores de la base de datos *Pedestrian Attribute*, utilizada en este Trabajo Fin de Máster, propusieron un enfoque alternativo para el reconocimiento de atributos. En él, en vez de tratar una imagen de forma independiente, se trata a la misma en base a la influencia del conjunto de sujetos con similares apariencias en su entorno.

Los resultados que demostraron utilizando la base de datos PETA fueron evaluados obteniendo el rendimiento con los siguientes sistemas: intersección de Kernel en SVM (ikSVM), MRF con Kernel Gaussiano (MRFg) y bosque aleatorio MRF (MRFr).

Para obtener el rendimiento, dividieron al conjunto de 19mil imágenes, asignando 9500 para entrenamiento, 1900 para verificación y 7600 para pruebas. Seleccionaron 35 atributos, de los cuales 15 son los más importantes en videovigilancia propuesto por expertos [31] [32], y el resto cubren todas las partes del cuerpo.

Los resultados, anotados en la siguiente tabla, demostraron que la detección precisa de atributos suaves puede lograrse con la vecindad inferida automáticamente.

Attribute	ikSVM	MRFg1	MRFg2	MRFr1	MRFr2
Age16-30	80.4	80.9	81.7	80.9	**83.8**
Age31-45	73.6	74.6	76.2	74.0	**78.8**
Age46-60	73.1	74.1	75.2	73.2	**76.4**
AgeAbove61	87.2	87.2	88.2	86.3	**89.0**
Backpack	66.7	67.1	67.1	67.0	**67.2**
CarryingOther	64.6	64.9	66.8	64.6	**68.0**
Casual lower	70.7	70.9	**71.6**	70.4	71.3
Casual upper	70.3	70.4	71.2	69.8	**71.3**
Formal lower	71.0	71.2	71.8	71.2	**71.9**
Formal upper	70.0	70.3	**70.4**	70.3	70.0
Hat	82.3	82.9	84.3	82.3	**86.7**
Jacket	67.7	68.3	**68.4**	68.1	67.9
Jeans	74.9	75.2	**76.1**	75.0	76.0
Leather shoes	78.9	80.1	80.9	79.1	**81.7**
Logo	**51.1**	**51.1**	**51.1**	**51.1**	50.7
Long hair	71.5	71.7	72.6	71.8	**72.8**
Male	79.7	80.3	80.9	80.6	**81.4**
MessengerBag	71.8	72.9	74.3	72.7	**75.5**
Muffler	88.0	88.3	89.5	86.5	**91.3**
No accessory	76.8	77.2	78.6	77.1	**80.0**
No carrying	70.4	70.6	**71.6**	70.6	71.5
Plaid	64.0	64.5	64.5	**65.0**	**65.0**
Plastic bag	74.9	74.9	75.5	73.9	**75.5**
Sandals	**50.3**	**50.3**	**50.3**	**50.3**	**50.3**
Shoes	70.6	71.0	72.5	70.8	**73.6**
Shorts	56.0	**56.5**	**56.5**	**56.5**	**56.5**
ShortSleeve	71.3	71.7	**71.8**	71.7	71.6
Skirt	64.0	64.0	64.0	64.0	**64.3**
Sneaker	67.5	68.1	69.0	68.2	**69.3**
Stripes	51.5	**52.3**	**52.3**	**52.3**	**52.3**
Sunglasses	**52.4**	**52.4**	**52.4**	51.8	51.7
Trousers	74.0	74.5	75.7	75.7	**76.5**
T-shirt	64.3	64.5	**64.6**	63.6	64.2
UpperOther	80.7	80.7	81.8	81.1	**83.9**
V-Neck	**51.1**	**51.1**	**51.1**	**51.1**	**51.1**
AVERAGE	69.5	69.9	70.6	69.7	**71.1**

Tabla 2.1 Rendimiento obtenido con la base de datos PETA [30]

2.5.2 Datasets soft biometrics

Actualmente existen más de 30 bases de datos publicas disponibles para la identificación de personas, sin embargo, solo unas pocas contienen los rasgos biométricos suaves etiquetados. La gran mayoría contiene información de la cara, donde existen más atributos a extraer.

A continuación, se describen algunas de las bases de datos más conocidas y utilizadas para el reconocimiento de personas utilizando estos atributos a lo largo de las últimas décadas.

- Faces in the Wild (LFW): Base de datos pública de imágenes de rostros creada y mantenida por investigadores de la Universidad de Massachusetts, contiene más de 13000 imágenes de 5749 personas que fueron detectadas y centradas por el detector facial Viola Jones [33], y recolectadas en su web. Entre las

personas etiquetadas, 1680 personas tienen dos o más fotos distintas. Sus creadores han sido galardonados con el premio Mark Everingham en octubre 2019 por su labor en la comunidad de Visión Artificial.

- Olivetti Research Laboratory (ORL): Base de datos pública formada por 400 imágenes en escala de grises de 40 individuos, 10 imágenes de cada uno de tamaño 112x92pixeles. Las imágenes presentan variaciones en las expresiones faciales como ojos abiertos o cerrados y algunas poseen accesorios como gafas. Además, las imágenes presentan cambios de iluminación.

- Yale: Base de datos pública de caras en escala de grises compuesta por 11 imágenes etiquetadas de cada uno de los 15 individuos que la forman. Cada imagen corresponde a una expresión facial o configuración diferente. [34]

- Pedestrian Intention Estimation (PIE): Base de datos para estudiar el comportamiento de 1842 viandantes. Con más de 6h de video, presenta información del comportamiento de las personas en el tráfico.

Figura 2.7: En orden, imágenes de las bases de datos ORL, Yale y PIE.

- Market-1501 dataset: formada por más de 32mil imágenes y 30 atributos diferentes, entre los que destacan 9 atributos binarios diferentes, contiene 4 grupos de diferentes edades e indica el color de las prendas superior e inferior de las personas

- DukeMTMC-reID dataset: formada con más de 36mil imágenes y 23 atributos etiquetados entre los que se encuentran 8 atributos binarios con colores de las prendas superiores e inferiores de las personas.

- Tunel multibiometrico:[35] base de datos de la Universidad Southampton diseñada con aeropuertos y otros entornos de alto rendimiento. Capaz de adquirir un conjunto de datos biométricos de una forma no intrusiva utilizando ocho cámaras sincronizadas para capturar la marcha de una persona a medida que atraviesa un túnel. También utiliza otras cámaras para capturar imágenes de la cara y oreja. Tiene una tasa de clasificación correcta de 99.6% utilizando la extracción de datos durante la marcha.

3.Entorno experimental

En este capítulo se describirá la base de datos utilizada para este proyecto y las razones por las cuales se ha elegido, así como el preprocesado realizado a estos datos. Además, se explica el diseño de las redes previamente entrenadas para el correcto desarrollo en el siguiente capítulo.

3.1 Análisis de la base de datos

El objetivo de este trabajo es comparar y valorar los resultados que proporciona las redes neuronales convolucionales para la identificación de rasgos biométricos suaves en la distancia como son los atributos corporales en una imagen con viandantes a distancia. Los resultados de dichas redes, nos proporcionará un valor de acierto o error que se valorará en el Capítulo Pruebas y Resultados.

Para el desarrollo de este Trabajo Fin de Máster se ha utilizado la base de datos *Pedestrian Attribute*, en adelante PETA, compuesta de imágenes a color de personas en la calle etiquetadas con multitud de atributos físicos de cada persona. Esta base de datos se puede obtener de forma libre en internet en [36].

El motivo por el que se ha escogido esta base de datos es porque PETA es una de las bases de datos etiquetadas más grandes que hay actualmente en el ámbito de viandantes de cuerpo entero. Esta base de datos, además, es muy diversa en variaciones de imágenes, escenarios y perspectiva.

Está formada por diez bases de datos diferentes conocidas en el campo de los rasgos biométricos suaves, que logran juntar un total de 19000 imágenes entre ellas. Entre todas ellas, incluye 8705 personas diferentes captadas desde diferentes ángulos y dispone de un tamaño de las imágenes que varía desde los 17x39 a 169x365 pixeles.

Figura 3.1: Base de datos PETA [30]

A continuación, se muestra en una tabla las características de estas bases de datos de forma individual.

Base de datos	**Núm. Imágenes**	**Ángulo cámara**	**Punto de vista**	**Iluminación**	**Resolución (pixeles)**	**Escena**
3DPeS	1012	alto	variable	variable	desde 31x100 a 236 x 178	exterior
CAVIAR4REID	1220	alrededor	variable	baja	desde 17x39 a 72x141	exterior
CUHK	4563	alto	variable	variable	80x160	interior
GRID	1275	variable	frontal y espalda	baja	desde 29x67 a 169x365	interior
i-LIDS	477	medio	espalda	alta	desde 32x76 a 115x294	exterior
MIT	888	alrededor	espalda	alta	64x128	exterior
PRID	1134	alto	perfil	baja	64x128	exterior
SARC3D	200	medio	variable	variable	desde 54x187 a 150x307	exterior
TownCentre	6967	medio	variable	medio	desde 44x109 a 148x332	exterior
VIPeR	1264	alrededor	variable	variable	48x128	exterior
Total = PETA	**19000**	**variable**	**variable**	**variable**	**variable**	**variable**

Tabla 3.1: Características base de datos PETA

Esta base de datos está etiquetada con 65 atributos, de los cuales 61 son binarios y 4 son multiclase. Estos últimos diferencian hasta once colores cada uno de ellos.

La distribución de un atributo binario se considera balanceado si el ratio entre la clase más grande y la más pequeña no es más de 20:1. En PETA, 31 de los 61 atributos binarios son balanceados.

Entre todos, cabe destacar atributos demográficos y de material o comportamiento; como son el género o edad de la persona, la apariencia física como puede ser el pelo, el tipo de ropa que llevan puesta o los accesorios.

3.2 Rasgos biométricos suaves seleccionados

Como se ha mencionado previamente, la base de datos PETA posee 65 atributos físicos etiquetados en cada imagen siendo binarios o multiclase.

En la siguiente tabla, se puede observar el total de los atributos que dispone PETA, siendo los cuatro últimos marcados con la M de multi-etiqueta los que tienen 11 colores diferenciados.

PETA DATASET ATTRIBUTES				
accessoryHeadphone	carryingUmbrella	footwearLeatherShoes	upperBodyNoSleeve	carryingSuitcase
personalLess15	lowerBodyCasual	upperBodyLogo	upperBodyPlaid	lowerBodySuits
personalLess30	upperBodyCasual	hairLong	carryingPlasticBags	accessorySunglasses
personalLess45	personalFemale	lowerBodyLongSkirt	footwearSandals	upperBodySweater
personalLess60	carryingFolder	upperBodyLongSleeve	footwearShoes	upperBodyThickStripes
personalLarger60	lowerBodyFormal	lowerBodyPlaid	hairShort	lowerBodyTrousers
carryingBabyBuggy	upperBodyFormal	lowerBodyThinStripes	lowerBodyShorts	upperBodyTshirt
carryingBackpack	accessoryHairBand	carryingLuggageCase	upperBodyShortSleeve	upperBodyOther
hairBald	accessoryHat	personalMale	lowerBodyShortSkirt	upperBodyVNeck
footwearBoots	lowerBodyHotPants	carryingMessengerBag	footwearSneaker	footwear (M x11)
lowerBodyCapri	upperBodyJacket	accessoryMuffler	footwearStocking	hair (M x11)
carryingOther	lowerBodyJeans	accessoryNothing	upperBodyThinStripes	lowerbody (M x11)
carryingShoppingTro	accessoryKerchief	carryingNothing	upperBodySuit	upperbody (M x11)

Tabla 3.2: Atributos base de datos PETA

De entre todos los atributos etiquetados que contiene la base de datos, se han elegido los más característicos de una persona a simple vista.

Para probar el reconocimiento de la red neuronal convolucional se han escogido todos los atributos multiclase que posee y, además, se ha convertido atributos binarios en multiclase, uniendo características para formar más atributos multiclase.

Debido a que no todos los atributos y sus tipos tienen el mismo número de imágenes, se ha hecho una selección previa dentro de los atributos seleccionados con mayor número de imágenes, puesto que las CNN necesitan gran cantidad de información para el entrenamiento, y de esta forma se tendrá un conjunto amplio de muestras para entrenar.

Dentro del conjunto de imágenes, se puede diferenciar una calidad de imagen muy variable entre todas ellas. Además, después de realizar el reajuste de tamaño, la calidad empeora. Este aspecto se valorará también en el capítulo de resultados.

Los atributos escogidos han sido divididos según la parte del cuerpo, llegando algunos a estar implicados en dos o incluso las tres secciones.

1) **Sección 1**: En esta división se encuentra generalmente centrada la cabeza del individuo y hombros. Algunas imágenes pueden contener parte del pecho. Para esta sección se han identificado los atributos y clases indicados en la siguiente tabla:

Atributo	Clases
Color de pelo	Pelo blanco
	Pelo gris
	Pelo marrón
	Pelo negro
	Pelo rubio
Edad	Menor de 15 años
	Comprendida entre 15 y 30 años
	Comprendida entre 30 y 45 años
	Comprendida entre 45 y 60 años
	Mayor de 60 años
Longitud de pelo	Pelo largo
	Pelo corto
	Calvo
Color prenda superior	Blanco
	Gris
	Negro
Género	Hombre
	Mujer

Tabla 3.3: Atributos seleccionados en sección 1

2) **Sección 2:** En esta división se puede observar de forma aproximada la distribución de los individuos desde el pecho hasta próximo a las rodillas. La siguiente tabla contiene la información de los atributos escogidos.

Atributo	Clases
Color prenda superior	Blanco
	Gris
	Negro
Color prenda inferior	Azul
	Gris
	Marrón
	Negro
Género	Hombre
	Mujer

Tabla 3.4: Atributos seleccionados en sección 2

3) **Sección 3.** Esta última división de la imagen comprende la parte inferior de las piernas y los pies. Los atributos escogidos se muestran en la siguiente tabla:

Atributo	Clases
Color prenda inferior	Azul
	Gris
	Marrón
	Negro
Genero	Hombre
	Mujer
Zapatos	Zapatillas
	Zapatos

Tabla 3.5: Atributos seleccionados en sección 3

Como se puede observar, no se han escogido todas las clases de ciertos atributos, dado que existe una gran desigualdad de imágenes entre algunas etiquetas y esto impediría el correcto entrenamiento de las redes neuronales convolucionales. Esto será analizado en el capítulo 5 mostrando los resultados dada una red entrenada con clases desbalanceadas y con las mismas clases balanceadas.

3.3 Redes neuronales convolucionales previamente entrenadas

De las múltiples redes neuronales de aprendizaje profundo que existen, en la primera parte de este trabajo se trabajarán las redes neuronales convolucionales empleando el método fine-tuning para obtener los resultados de clasificación de los rasgos biométricos suaves seleccionados de la base de datos PETA.

La capa de salida realizará la clasificación de la etiqueta correspondiente. De esta forma, la propia red convolucional realiza las tareas de extracción de características y clasificación.

Las redes previamente entrenadas en la realización de este trabajo han sido AlexNet y GoogLeNet que se detallarán a continuación.

AlexNet

Alexnet es una red neuronal convolucional creada por Alex Krizhevskycon y presentada en el desafío de Reconocimiento Visual a Gran Escala de ImageNet, ILSVRC, en septiembre de 2012 por Ilya Sutskever y Geoffrey E. Hinton [37].Consiguió el primer puesto con una tasa de error de un 15.3%, más de 10 puntos de porcentaje de diferencia sobre el subcampeón.

Inicialmente se dijo que la profundidad del modelo era esencial para su rendimiento, siendo computacionalmente costoso pero factible, debido a la utilización de GPU durante el entrenamiento.

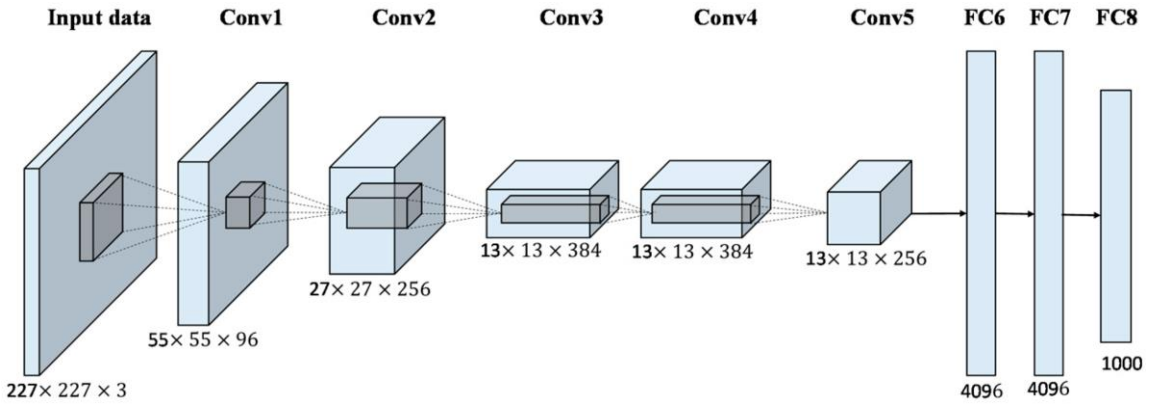

Figura 3.2: Red Alexnet

Está compuesta por un total de 25 capas, entre las que se incluyen 5 capas convolucionales con la función de activación ReLU, 3 capas *max-pooling*, 2 capas de normalización, 3 capas *fully connected* + ReLU y una capa *softmax*. En total suma 60 millones de parámetros y 650 mil neuronas lo que suponía un problema de sobreajuste u *overfitting*.

El sobreajuste se produce cuando no hay suficientes muestras con las que entrenar y por tanto aprender la red. Por eso se emplearon dos métodos para reducir este sobreajuste en su entrenamiento:

- **Aumento de datos o *Data Augmentation***: Los autores utilizaron la transformación de imágenes para hacer que sus datos fueran más variados y aumentaran en número. En concreto, realizaron traslaciones en las imágenes y reflexiones horizontales, lo que aumentó el conjunto de entrenamiento en un factor de 2048.

  Además, realizaron el Análisis de Componentes Principales (PCA) en los valores de pixeles RGB cambiando sus intensidades, lo que redujo el error en más de 1%.

- **Dropout**: Alexnet tiene 2 capas de este tipo. Es una técnica que consiste en "apagar" ciertas neuronas en la capa oculta, de tal forma que cada iteración usa una muestra diferente de los parámetros del modelo, lo que obliga a cada neurona a tener características más robustas que pueden ser usadas por otras neuronas aleatorias. Sin embargo, esta técnica también aumenta el tiempo de entrenamiento necesario para la convergencia del modelo.

La entrada de imágenes a la red debe tener un tamaño de 227x227x3 pixeles, aunque originalmente fue entrenada con imágenes de 224x224x3 pero un *padding* fue añadido en los bordes. Esta red ha sido entrenada para diferenciar entre 1000 objetos diferentes.

GoogLeNet

GoogLeNet es una red neuronal convolucional ganadora del reto ILSVRC en 2014. Como su nombre indica, fue creada por Google y le rinde homenaje a la red LeNet del profesor Yan LeCun´s. También es conocida por llamarse red *Inception* V1. [38]

Con esta red se introduce un nuevo nivel de organización, la capa *Inception*. La idea de esta capa es cubrir un área más grande, pero también mantener una resolución clara para la información más pequeña de las imágenes.

Se trata de convolucionar en paralelo diferentes tamaños de filtros, desde el detalle más preciso (1x1pixel) hasta uno más grande como puede ser (5x5 pixeles). Con esta técnica una serie de filtros con diferentes tamaños manejarán mejor las escalas de los múltiples objetos de la imagen.

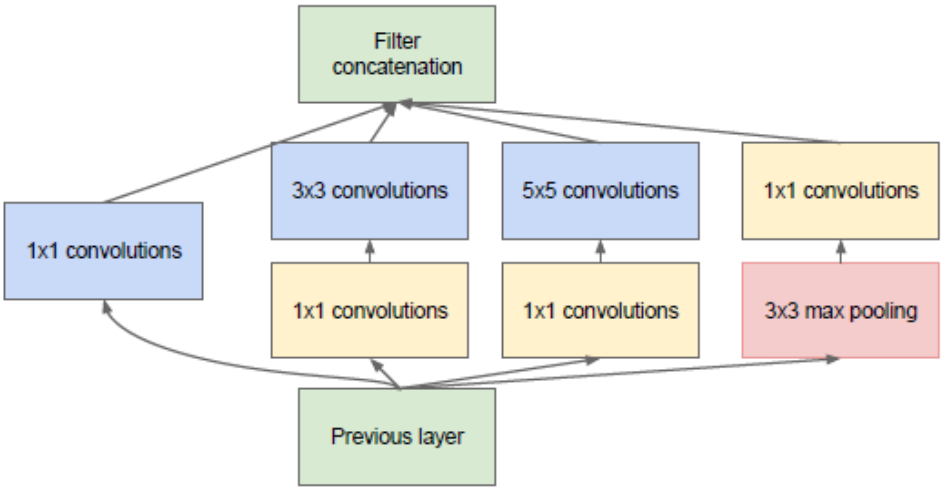

Figura 3.3: Módulo Inception de la red GoogLeNet

GoogLeNet usa nueve módulos Inception. El problema es que más parámetros también significa que el modelo es más propenso a sobreajustarse. Para evitar este problema, la técnica cuello de botella fue aplicada.

Esta técnica consiste en insertar entre medias una capa con menos neuronas que la capa previa y posterior. Con ello, esta red reduce el número de canales, pero al mismo tiempo, permite que sea profunda y represente muchos mapas de características.

Utilizando las técnicas de cuello de botella, se puede reconstruir el módulo Inception con la no-linealidad y menos parámetros. Además, se utiliza una capa *max-pooling* para resumir el contenido de la capa previa.

Esta red tiene un total tiene 144 capas. En la siguiente tabla se puede observar la arquitectura de los módulos Inception.

type	**patch size/ stride**	**output size**	**depth**	#1×1	#3×3 **reduce**	#3×3	#5×5 **reduce**	#5×5	**pool proj**	**params**	**ops**
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Tabla 3.6: Arquitectura Inception de la red GoogLeNet

4.Diseño y desarrollo del Sistema

En este capítulo se describirá en detalle cómo se ha llevado a cabo la implementación de las redes neuronales convolucionales. Es decir, se describirá la parte más importante del proyecto, sobre la que luego se evaluarán los resultados. El conjunto del proyecto se ha desarrollado en Matlab.

El objetivo principal es conseguir una clasificación de los rasgos biométricos suaves seleccionados dada una imagen de un viandante en la lejanía de forma automática.

En primer lugar, se evaluarán las redes previamente entrenadas Alexnet y GoogLeNet realizando la técnica aprendizaje por transferencia, para más tarde crear una red neuronal convolucional desde cero.

La finalidad es poder comparar los resultados entre estas redes galardonadas en ILSVRC por tener los mejores resultados en clasificación, pero para las que han sido entrenadas para clasificar otros objetos, con una red entrenada desde cero para clasificar únicamente este tipo de atributos.

4.1 Preprocesado de la base de datos

El primer paso que se ha realizado es el procesado de la base de datos para introducir las imágenes en nuestras CNNs.

La base de datos utilizada, PETA, fue elegida por la gran cantidad de rasgos biométricos suaves de los que disponía. Formada por 19000 imágenes y etiquetada con un total de 65 atributos.

En primer lugar, para poder trabajar con todas las imágenes, sin importar el origen de la base de datos individual, se ha etiquetado el nombre de cada imagen renombrando de 0 a 19000 sin distinguir cuando el mismo individuo se repite. Pues hay muchas imágenes de las mismas personas, tomadas desde diferentes distancias y/o ángulos. En concreto entre las 19000 imágenes, hay 8705 personas diferentes. De esta forma se tendrán más imágenes para el entrenamiento y la validación en el proceso.

Figura 4.1: Imágenes de la misma persona en PETA desde diferentes ángulos con calidades y tamaños variados 295x101, 270x100, 288x90x3 y 264x120x3

4.1.1 Segmentado de las imágenes

Debido al movimiento del cuerpo de las personas y el ángulo de las fotos tomadas, algunos atributos son más difíciles de captar. Además, algunos atributos se encuentran localizados en partes concretas del cuerpo, como el color de pelo que se puede encontrar en la cabeza de las personas. Por este motivo, se ha realizado una división de cada imagen en tres partes iguales.

Todas las imágenes originales tienen forma rectangular variando sus tamaños entre 17x39 pixeles a 169x365 pixeles. Tras analizar previamente las imágenes, se observó que prácticamente en su totalidad, las imágenes tienen centrada a la persona y por tanto se puede localizar ciertos atributos que deseamos clasificar en áreas concretas.

Cada imagen original ha sido dividida en tres, separando cabeza-hombros, torso-cadera y piernas-pies. La división se ha realizado simétricamente, tomando el total de altura en pixeles de cada imagen y dividiendo en tres secciones. De ahora en adelante, a esta subdivisión de cada imagen se nombrará en este documento como sección 1, 2 y 3 respectivamente.

Figura 4.2: Segmentado de imágenes

4.1.2 Reajuste de tamaño

Por último, el total de las imágenes segmentadas en las tres secciones han sido reajustadas en tamaño para poder introducirlas en las redes de Alexnet y GoogLeNet. Ambas redes admiten las imágenes en color, RGB.

El tamaño de las imágenes de entrada que necesita la red de Alexnet es de 227x227x3 pixeles, mientras que la red GoogLeNet necesita un tamaño de imágenes de entrada de 224x244x3.

A continuación, se muestra un ejemplo de los cambios aplicados a las imágenes tras el reajuste de tamaño indicando en cada sección los atributos seleccionados para analizar en este trabajo.

Figura 4.3: Reajuste de tamaño para red Alexnet 227x227 pixeles

Algunos de los atributos seleccionados para su clasificación se pueden observar en más de una sección. Además, se ha escogido el rasgo biométrico suave "género" para ver si la red es capaz de identificar el género de la persona dándole no solo la cara para analizar, sino también el cuerpo o las piernas. Este análisis sería, en gran parte de las ocasiones, muy difícil de identificar hasta por un ser humano. Este aspecto será destacable en los resultados obtenidos.

Para cada red convolucional, se introduce cada atributo individual seleccionando exclusivamente la parte o partes donde se encuentra localizado el atributo elegido en cada momento. De esta forma se puede analizar los mejores resultados obtenidos entre todas las pruebas.

No todas las imágenes disponibles de la base de datos se utilizan para todos los atributos, un ejemplo de esto se puede ver en la Figura 4.3 que es apta para todos los atributos menos el color de la prenda superior porque el color rojo no ha sido seleccionado al no tener suficientes imágenes para entrenar con ese color.

El conjunto de imágenes utilizadas para cada atributo se ha dividido en entrenamiento, validación y pruebas. Para almacenar estos conjuntos de imágenes se ha utilizado la función de Matlab, ImageDatastore. Esta función de Matlab además de etiquetar cada imagen con el nombre de la carpeta en la que se encuentre, evitando un fichero de texto que lo enlace, permite almacenar gran cantidad de datos en un objeto de tipo

ImageDatastore, incluidos los datos que no caben en la memoria, y leer de manera eficiente el conjunto de imágenes durante el entrenamiento de la red.

Aun cuando el tamaño de nuestra base de datos no es comparable a las utilizadas para entrenar otras redes, supone gran cantidad de almacenamiento necesario al tener que multiplicar esas imágenes por el número de atributos y secciones a utilizar. Por ese motivo, almacenar los conjuntos de imágenes en este tipo de objeto fue la mejor opción.

4.2 Transfer learning Alexnet.

El principal motivo por el que se realiza el aprendizaje por transferencia sobre redes previamente entenadas es porque se puede solventar el problema de clasificación de imágenes dado un conjunto pequeño de datos.

El primer paso fue aplicar esta técnica sobre la red de Alexnet. Alexnet ha sido entrenada con más de un millón de imágenes para clasificarlas entre 1000 categorías diferentes, y se ha modificado para que sea capaz de clasificar las imágenes de viandantes entre las clases de los atributos seleccionados, variando el número de clases por atributo desde 2 a 5 tipos o etiquetas.

4.2.1 Capas de la red Alexnet

Es necesario cargar en el entorno de trabajo la red completa de Alexnet gracias a la herramienta de Matlab *Deep Learning Toolbox Model for Alexnet Network*. Tras el análisis de la red, se extraen las tres últimas capas: la capa totalmente conectada con las 1000 salidas que clasifica la red original, la capa *softmax* que es la función de transferencia, y la capa de salida que clasifica las categorías.

Estas tres capas, se reemplazan por los mismos tres tipos de capas, pero modificando los parámetros que se indican a continuación para que sea capaz de clasificar entre las clases de cada atributo. Comienza con este reemplazo de capas el proceso de aprendizaje por transferencia.

Para cada análisis de rasgo biométrico suave, lo primero fue extraer el número de clases o etiquetas del atributo a analizar en cada momento. Y se realizaron diferentes pruebas de capas finales a la red para ver cual obtenía los mejores resultados.

Entre estas pruebas se tuvo en cuenta que una red de clasificación, dado que la respuesta que queremos es categórica, debe tener como mínimo una capa totalmente conectada al final, seguido de una función de transferencia y finalmente la capa de salida por clasificación. Que equivalen a las tres capas reemplazadas de Alexnet.

La nueva capa totalmente conectada tendrá el parámetro número de clases para cada atributo a analizar, este número puede variar de 2 a 5 clases diferentes. Además, se ha configurado el factor de la tasa de aprendizaje para los pesos en 20, lo que implica que la red será veinte veces más rápida en aprender los pesos en esta capa. Siendo la velocidad de aprendizaje de los pesos global determinada por el software según las opciones de

entrenamiento. De la misma forma, se ha configurado en esta capa el factor de la tasa de aprendizaje para los sesgos en 20.

Tras la capa totalmente conectada, se añade la función de transferencia no lineal elegida, *softmax*, que asigna las probabilidades a la imagen en análisis entre las clases a clasificar.

Por último, se añade la capa de salida o clasificación que dará el resultado de la clase que tenga la probabilidad más alta.

	Alexnet	**Aprendizaje por transferencia**
Capa 23	`fully connected layer ( 1000)`	`fully connected layer ( num_clases, 'WeightLearnRateFactor',20, 'BiasLearnRateFactor',20)`
Capa 24	`Softmax layer`	`Softmax layer`
Capa 25	`Classification Output Layer`	`Classification Output Layer`

Tabla 4.1: Aprendizaje por transferencia en la red Alexnet

4.2.2 Opciones de entrenamiento de la red Alexnet

Para evitar que la red sufra sobreajuste durante el entrenamiento y validación, se ha utilizado la técnica de aumento de datos realizando la reflexión y traslación aleatoria de un rango de 20 pixeles.

Para las opciones de entrenamiento se ha fijado en primer lugar el factor de aprendizaje inicial que afectará a todas las capas. De nuevo, se ha probado a entrenar la red con algunos parámetros modificando sus valores y realizando una balanza entre la mejora de los resultados y el tiempo que tarda en entrenar.

Para esta red, se ha probado con valores de 0.001, 0.0001 y 0.00001. Como este último aumentaba en exceso el tiempo de entrenamiento y no mejoraba los resultados se decidió fijar el factor de aprendizaje inicial en 0.0001.

El número de épocas fijado para entrenar esta red ha sido 10 épocas tras ver que los resultados dejaban de mejorar prácticamente entre las épocas 6 y 7 con cada atributo. Las épocas de diferencia hasta 10 no produce sobreajuste y simplemente permite que la red se estabilice obteniendo el mejor resultado. En cada comienzo de época, las imágenes se mezclarán para que no influya el orden en su entrenamiento. Esta configuración se podrá ver mejor en el Capítulo 5 en la sección de pruebas.

Para validar el conjunto de imágenes de entrenamiento, se ha utilizado el conjunto de imágenes dividido para validación con una frecuencia de validación cada 30 iteraciones.

4.3 Transfer Learning GoogLeNet

La segunda red escogida, GoogLeNet, entrenada con más de un millón de imágenes etiquetadas y con alta resolución de la base de datos ImageNet, es capaz de clasificar una imagen en 1000 categorías diferentes. Con ella, se realiza de nuevo un aprendizaje por transferencia sobre la red para que sea capaz de diferenciar los rasgos biométricos suaves seleccionados de la base de datos PETA.

Como pasó con la red Alexnet, para cargar la red GoogLeNet en nuestro entorno de trabajo, será necesario utilizar la herramienta de Matlab *Deep Learning Toolbox Model for GoogLeNet.*

Tras el análisis inicial de la red, se reajusta el conjunto de imágenes en las bases de datos de entrenamiento, validación y entrenamiento porque la entrada a la red deben ser imágenes de tamaño 224x224x3 pixeles.

4.3.1 Capas de la red GoogLeNet

Para este aprendizaje por transferencia se han eliminado las tres últimas capas de la red, pues son las que clasifican las imágenes entre las 1000 categorías con las que ha sido entrenada. Contiene las mismas últimas tres capas que la red Alexnet, son: una capa totalmente conectada con las 1000 categorías a clasificar, una capa *softmax* que asigna las probabilidades dentro del problema multiclase y, por último, la capa de salida que proporciona la clasificación.

Se ha sustituido la capa totalmente conectada por otra, pero con el número de salidas equivalente al número de clases del atributo en estudio en cada momento. Además, se ha configurado el factor de la tasa de aprendizaje para los pesos en 10 y el factor de la tasa de aprendizaje para los sesgos en 10, lo que implica que la red será diez veces más rápida en aprender los pesos y sesgos en esta capa.

Se ha añadido la capa *softmax* para que indique las probabilidades que tiene cada imagen entre las clases a clasificar. Y, por último, se ha finalizado la red con la capa de salida de tipo clasificación para que determine la clase de la imagen tomando el máximo valor entre las probabilidades obtenidas hasta ese momento.

Para acelerar aún más el entrenamiento de esta red, se han fijado los pesos de las diez primeras capas de la red GoogLeNet a cero porque son las capas que forman el tronco principal, justo antes de comenzar el primer módulo Inception. De tal forma que estos parámetros no se actualizan en esas capas porque no es necesario calcular sus gradientes de nuevo. Además de acelerar, esta técnica puede evitar que esas capas sufran sobreajuste con el conjunto de imágenes nuevas.

	GoogLeNet	Aprendizaje por transferencia
Capa 142	`fully connected layer ( 1000)`	`fully connected layer ( num_clases, 'WeightLearnRateFactor',10, 'BiasLearnRateFactor',10)`
Capa 143	`Softmax layer`	`Softmax layer`
Capa 143	`Classification output`	`Classification output`

Tabla 4.2: Aprendizaje por transferencia en la red GoogLeNet

4.3.2 Opciones de entrenamiento para GoogLeNet

Para evitar que la red sufra sobreajuste durante el entrenamiento y validación, se ha utilizado la técnica de aumento de datos realizando la reflexión y traslación aleatoria de un rango de 30 pixeles. Además, se han escalado aleatoriamente hasta un 10% horizontal y verticalmente las imágenes. Esto permite tener más imágenes para entrenar pues a la red se le presenta diferentes.

Para las opciones de entrenamiento se ha fijado en primer lugar el factor de aprendizaje inicial que afectará a las capas que no se ha fijado a cero. De nuevo, se ha probado a entrenar la red con algunos atributos con diferentes factores, disminuyendo su valor y realizando una balanza entre la mejora de los resultados y el tiempo que tarda en entrenar.

Para esta red, se ha probado con valores de 0.003, 0.0003 y 0.00003. Como este último aumentaba en exceso el tiempo de entrenamiento y no se mejoraba más que un máximo de medio punto porcentual en algún atributo, se decidió fijar el factor de aprendizaje inicial en 0.0003.

Este factor de aprendizaje inicial más lento junto con los factores de aprendizaje de pesos y sesgos fijados en la capa nueva totalmente conectada produce un aprendizaje más rápido en las nuevas capas añadidas tras el aprendizaje por transferencia, un aprendizaje más lento en las capas intermedias y ningún aprendizaje en las primeras capas donde se ha fijado la tasa a cero. Combinando estas tres técnicas, se consigue el mejor rendimiento-tiempo de entrenamiento de la red.

El número de épocas fijado para entrenar esta red ha sido 10 épocas tras ver que los resultados dejaban de mejorar prácticamente entre las épocas 6 y 8 con cada atributo. Este número de épocas definido no produce que sobreajuste el modelo. En cada comienzo de época, las imágenes se mezclarán para que no influya el orden en su entrenamiento.

Para validar el conjunto de imágenes de entrenamiento, se ha utilizado el conjunto de imágenes dividido para validación con una frecuencia de validación cada 30 iteraciones.

4.4 Configuración red CNN propia

Una vez conocido el funcionamiento de las redes neuronales y haber realizado el proceso de aprendizaje de transferencia con dos redes comúnmente conocidas, se crea desde cero una red propia definiendo las capas y las opciones de entrenamiento.

Para esta red, se han tomado la misma base de datos, PETA, con el mismo conjunto de imágenes para el entrenamiento, validación y prueba que para la red Alexnet, pues se ha definido una capa de entrada para imágenes de 227x227x3 pixeles.

4.4.1 Capas de la red propia

En primer lugar, se definen las capas que forman la red neuronal convolucional para la extracción de rasgos biométricos suaves. La relevancia principal la tomarán las capas convolucionales porque son las que permitirán a la red extraer los atributos de forma automática.

Durante el proceso de creación de la nueva red, se fueron realizando diferentes arquitecturas añadiendo capas convolucionales con diferentes tamaños y número de filtros. Las pruebas se realizaban con un atributo de cada sección: color de pelo para la sección 1, color de la parte superior para la sección 2 y color de la parte inferior para la sección 3.

En los tres tipos de pruebas, se ha utilizado el atributo que se diferencia de sus clases por el color, para que en las tres secciones la característica a extraer sea el color, dado que el único atributo en común con las tres secciones es el género, pero no tiene la misma dificultad extraer este atributo según el tramo del cuerpo.

Tras el análisis del conjunto de imágenes y dado el segmentado de las mismas con los atributos que se quieren analizar, finalmente se decidió por utilizar la red con tres capas convolucionales con pocos y pequeños filtros para no perder detalle en secciones que ya de por sí se encuentran acotadas:

- Conv_1: configurada con 8 filtros de tamaño 5x5. Y con *padding* de 4 pixeles en cada borde de la imagen. Esto se ha hecho para que inicialmente no pierda detalle de los extremos de las imágenes, a pesar de tener una base de datos con el individuo principalmente centrado, se pretende que esta red tenga validez con otras imágenes que no sean las de PETA y el individuo pueda encontrarse en una esquina.
- Conv_2: configurada con 8 filtros de tamaño 3x3, y con un *padding* de 1 píxel en cada borde de la imagen. A medida que avanza la red, se quiere extraer más características fijándose en detenimiento, por eso disminuye el tamaño de los filtros. Y se disminuye el *padding* con respecto a la primera capa de convolución para que tome más relevancia la parte central.
- Conv_3: de nuevo, configurada con 8 filtros de tamaño 3x3 y con un *padding* de 1 píxel en cada borde.

En las capas convolucionales, el *stride* es siempre de uno en uno pues no se quiere perder detalle por los atributos tan variados que se disponen. Siendo muy difícil identificar por ejemplo el rango de edad de una persona.

A continuación de las dos primeras capas convolucionales, se ha aplicado la normalización de las neuronas con la capa *batchNormalization*, se ha utilizado la función de activación ReLu para eliminar los datos negativos, y se ha reducido a la mitad la información utilizando una capa *max-pooling*.

Las capas *max-pooling* han sido configuradas con un tamaño de pool 2x2 que tomará el máximo valor de ese conjunto de pixeles reduciendo así el número de activaciones a la mitad. Además, se ha configurado un *stride* de 2, que implica el número de pixeles que se desplaza el filtro para cada convolución. Este *stride* se ha aplicado así porque si es menor que las dimensiones del pool, las regiones se solapan.

Para la última capa convolucional, conv_3, se ha aplicado de nuevo la capa *batchNormalization* para normalizar los valores de las neuronas y se ha utilizado la función de activación ReLu para eliminar los datos negativos.

El conjunto de las tres capas convolucionales permite que nuestra red tome los valores más importantes para extraer las características más relevantes y que se reduzca el número de parámetros para que la red no crezca exageradamente y su cómputo no sea muy elevado.

Tras las capas de convolución, finalmente está la capa totalmente conectada con el parámetro número de clases. Este parámetro es obtenido de forma automática al inicio del sistema según el atributo que esté analizando la red.

La capa totalmente conectada asocia las transformaciones hechas a las imágenes en categorías, siendo en nuestro caso las clases definidas.

Para finalizar, se coloca la función de activación de la capa totalmente conectada, *softmax*, que transforma las salidas en forma de probabilidades, de tal manera que el sumatorio de todas las probabilidades de las salidas de 1. Se ha utilizado la función de activación *softmax* en vez de sigmoide porque la mayoría de los atributos a extraer son multiclase.

Por último, se coloca la capa de clasificación, *Classification Output*, que calcula la pérdida de entropía cruzada en problemas multiclase donde las clases son mutuamente excluyentes. El número de neuronas de la cada de salida será el número de clases que posea cada atributo.

Estas tres últimas capas, formarán el decisor de nuestra red, asignando a la imagen de entrada la etiqueta o clase que posea el porcentaje más alto.

A continuación, se puede observar el conjunto de capas que forman la red con los parámetros de cada una.

Número de capa	Nombre de capa	Parámetros
1	Image Input	227x227x3
2	Convolution_1	8 5x5 convolutions, stride [1 1], padding [4 4 4 4]
3	Batch Normalization	Batch normalization
4	ReLU	ReLU
5	Max Pooling	2x2 max pooling, stride [2 2], padding [0 0 0 0]
6	Convolution_2	8 3x3 convolutions, stride [1 1], padding [1 1 1 1]
7	Batch Normalization	Batch normalization
8	ReLU	ReLU
9	Max Pooling	2x2 max pooling, stride [2 2], padding [0 0 0 0]
10	Convolution_3	8 3x3 convolutions, stride [1 1], padding [1 1 1 1]
11	Batch Normalization	Batch normalization
12	ReLU	ReLU
13	Fully Connected	num_class fully connected layer
14	Softmax	softmax
15	Classifcation Output	crossentropyex

Tabla 4.3: Capas de la CNN propia

4.4.2 Opciones de entrenamiento para la red propia

Como con las redes previamente entrenadas, para obtener las opciones de entrenamiento optimas, se ha utilizado el proceso prueba-error configurando los parámetros según los resultados que se iban obteniendo y utilizando los mismos atributos en cada sección que para la creación de las capas de la red.

En todo momento se ha utilizado *Stochastic Gradient Descent with Momentum* que es el método mejorado de SGD. Este método ayuda a acelerar el vector de gradientes en la dirección correcta, lo que conduce a una convergencia más rápida. El valor SGD *Momentum* fijado ha sido 0.9 que es el utilizado por defecto en este parámetro.

Otro parámetro en las opciones de entrenamiento, y el cual ha tenido la mayor relevancia en este proceso junto con el número de épocas, es el factor de aprendizaje inicial. Este factor debe ser un valor escalar positivo. Si el valor es muy pequeño, el entrenamiento tarda mucho en completar, mientras que si el valor es grande el entrenamiento puede converger antes de tiempo y el resultado no sería óptimo.

Para el factor de Aprendizaje Inicial, se fue probando de forma exponencial desde 0.01, 0.001, 0.0001 hasta 0.00001. En esta red, dado que no tiene mucha complejidad, el aumento de tiempo no era tan relevante comparada con la mejora en los resultados que se iban obteniendo. Por tanto, se fijó este parámetro en 0.00001.

El segundo parámetro de mayor importancia fue el máximo número de épocas. Cada época consiste en la ejecución completa del algoritmo de entrenamiento sobre las imágenes de entrenamiento. A cada época, la red va aprendiendo y mejorando sus pesos y sesgos, lo que permite tener mejores resultados.

Con este parámetro ocurre similar al factor de Aprendizaje inicial, a mayor número de épocas, mayor es el tiempo de ejecución en la red. Por ese motivo, es bueno representar el proceso de entrenamiento con épocas-precisión, además de épocas-función de perdidas, pues es muy visual el momento en el que la red deja de mejorar, y así poder fijar el número de épocas en ese momento.

Para esta red, dado que no ha sido entrenada previamente como las Alexnet o GoogLeNet, cuanto más entrenamiento realice la red, mejores resultados se pueden obtener. Pero el aumento de épocas en el entrenamiento también puede producir que el modelo llegue a sobreajustarse. Se ha fijado en 30 épocas todos los entrenamientos de esta red, para que alcance el máximo rendimiento posible en ese proceso. Cada inicio de época, se mezclarán las imágenes, para que no influya el orden de estas.

Para realizar la validación de la red se ha utilizado el conjunto de imágenes de validación que corresponden al 25,5% de las imágenes del atributo a analizar. Para este proceso se ha ajustado la frecuencia de validación de la red en 30 iteraciones.

5. Integración, pruebas y resultados

En este capítulo se mostrarán los resultados obtenidos en la clasificación de atributos seleccionados de la base de datos PETA para las dos redes previamente entrenadas modificadas con fine-tuning y para la red propia. Para obtener estos resultados se ha realizado un conjunto de pruebas previo para escoger el valor de los parámetros de entrenamiento.

Los resultados se han obtenido con el conjunto de imágenes de pruebas sobre las redes entrenadas con las imágenes de entrenamiento y validación.

5.1 División dataset PETA

Para los modelos de aprendizaje de máquina, los algoritmos necesitan aprender a base de un conjunto de datos etiquetados. Pero para conocer si el modelo funciona correctamente tras el entrenamiento, es necesario medir los resultados con un conjunto de datos que no han sido manejados por la red previamente.

Por este motivo, tras la división de las imágenes de los viandantes en tres secciones y división de las imágenes según sus atributos y clases en diferentes carpetas, se divide el conjunto de imágenes en tres, entrenamiento, validación y test.

- Datos de entrenamiento: como indica su nombre, son los datos aportados a la red para entrenar el modelo. En cada época, el modelo será entrenado una y otra vez con el mismo conjunto de imágenes para calcular el gradiente, actualizar los pesos y sesgos de la red.
- Datos de validación: es el conjunto de imágenes usado para validar el modelo durante el entrenamiento. El error en los datos de validación se supervisa durante el proceso de entrenamiento. El error de validación normalmente disminuye durante la fase inicial del entrenamiento, al igual que el error de los datos de entrenamiento. Sin embargo, cuando la red comienza a sobreajustar los datos, el error en el conjunto de imágenes de validación comienza a aumentar. Los pesos y sesgos de la red se guardan cuando alcanza el error mínimo del conjunto de validación.

- Datos de prueba: es el conjunto de imágenes que no se utilizan durante el entrenamiento, sino una vez se tiene la red entrenada, para comparar entre los diferentes modelos de redes neuronales convolucionales.

Como el número de imágenes no estaba balanceado entre atributos, se ha dividido cada atributo inicialmente según los siguientes porcentajes:

Inicialmente se generado un aleatorio de cada etiqueta y se ha seleccionado un 85% de los datos para entrenamiento y validación. El 15% restante se ha asignado como conjunto de datos de prueba.

Después, del 85% seleccionado, se ha separado un 70% para entrenamiento y 30 % para el conjunto de validación. En el siguiente gráfico, se pueden observar los porcentajes finales de una forma más clara.

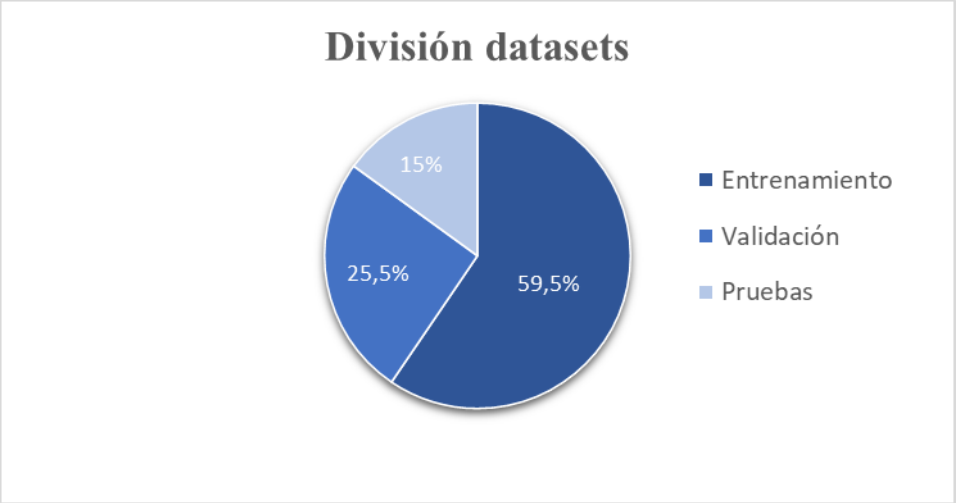

Figura 5.1: Porcentajes división dataset

Para cada atributo, esta división inicial se ha guardado, para así poder comparar los diferentes modelos de CNNs utilizando las imágenes exactas en cada uno.

A continuación, se puede observar en unas tablas el número de imágenes de cada atributo y sus clases con la división realizada en los tres diferentes datasets: entrenamiento, validación y pruebas.

SECCIÓN 1						
Atributo	**Clases**	**Número imágenes por clase**	**Dataset Entrenamiento**	**Dataset Validación**	**Dataset Test**	**Total**
Color de pelo	Pelo blanco	809	482	206	121	18346
	Pelo gris	1578	939	402	237	
	Pelo marrón	3986	2372	1016	598	
	Pelo negro	11491	6837	2930	1724	
	Pelo rubio	482	287	123	72	
Edad	Menor de 15 años	170	102	43	25	18972
	Comprendida entre 15 y 30 años	9441	5618	2407	1416	
	Comprendida entre 30 y 45 años	6246	3716	1593	937	
	Comprendida entre 45 y 60 años	1943	1156	496	291	
	Mayor de 60 años	1172	697	299	176	
Longitud de pelo	Pelo largo	4513	2685	1151	677	18757
	Pelo corto	13854	8243	3533	2078	
	Calvo	390	232	100	58	
Color prenda superior	Blanco	2840	1690	724	426	14502
	Gris	3130	1863	798	469	
	Negro	8532	5076	2176	1280	
Género	Hombre	10419	6199	2657	1563	18997
	Mujer	8578	5104	2187	1287	

Tabla 5.1: División de datasets sección 1

SECCIÓN 2						
Atributo	**Clases**	**Número imágenes por clase**	**Dataset Entrenamiento**	**Dataset Validación**	**Dataset Test**	**Total**
Color prenda superior	Blanco	2840	1690	724	426	14502
	Gris	3130	1863	798	469	
	Negro	8532	5076	2176	1280	
Color prenda inferior	Azul	3461	2059	883	519	17857
	Gris	4432	2637	1130	665	
	Marrón	806	479	206	121	
	Negro	9158	5449	2335	1374	
Género	Hombre	10419	6199	2657	1563	18997
	Mujer	8578	5104	2187	1287	

Tabla 5.2: División de datasets sección 2

SECCIÓN 3						
Atributo	**Clases**	**Número imágenes por clase**	**Dataset Entrenamiento**	**Dataset Validación**	**Dataset Test**	**Total**
Color prenda inferior	Azul	3461	2059	883	519	17857
	Gris	4432	2637	1130	665	
	Marrón	806	479	206	121	
	Negro	9158	5449	2335	1374	
Genero	Hombre	10419	6199	2657	1563	18997
	Mujer	8578	5104	2187	1287	
Zapatos	Zapatillas	4101	2440	1046	615	11000
	Zapatos	6899	4105	1759	1035	

Tabla 5.3: División de datasets sección 3

Como se puede observar, algunas clases se encuentran muy desbalanceadas dentro del mismo atributo, lo que puede perjudicar al entrenamiento de la red al tener más imágenes de una clase que otra. Por este motivo, se ha realizado la prueba de balancear las clases, igualando el número de imágenes de entrenamiento y validación de cada una a la clase con el mínimo número de imágenes. Esto se mostrará en las tablas de resultados en los siguientes apartados de este capítulo.

Los resultados se van a medir dado el rendimiento o *accuracy* obtenido con el conjunto de imágenes prueba en cada red, que se calcula como el número de atributos que han sido clasificados en su clase correcta respecto al número de imágenes de entrenamiento que se dispone. Esto nos proporciona un porcentaje que será el valor para comparar entre las diferentes redes.

5.2 Entorno de pruebas

En este apartado se van a analizar las pruebas realizadas entrenando las diferentes redes antes de decidir el número de épocas por red y conjunto de datos.

Para realizar estas pruebas, se han realizado diferentes entrenamientos con diferentes números de épocas. Se comenzó desde un mínimo, 6 épocas, considerando que algunas de las imágenes son complejas por su borrosidad y menos épocas no serían suficientes.

Para estas pruebas se han utilizado siempre el mismo atributo para cada sección: color de pelo para la sección 1, color de la parte superior para la sección 2 y color de la parte inferior para la sección 3. En las tres secciones se ha tomado un color como referencia porque como indicó Paisitkriangkrai et al [25], las redes previamente entrenadas con la base de datos ImageNet no considera el color como atributo importante.

Y dado que las dos redes analizadas con aprendizaje por transferencia en este trabajo han sido entrenadas y galardonadas por esta base de datos, se quiere observar el nivel de rendimiento que obtiene con el color. Destacando que del total de 11 atributos que se van a analizar entre las tres secciones, 5 de ellos hacen referencia al color de una parte el cuerpo.

Debido a que las gráficas que aparecerán a continuación son demasiado grandes, esta será la leyenda de colores utilizadas en todas ellas:

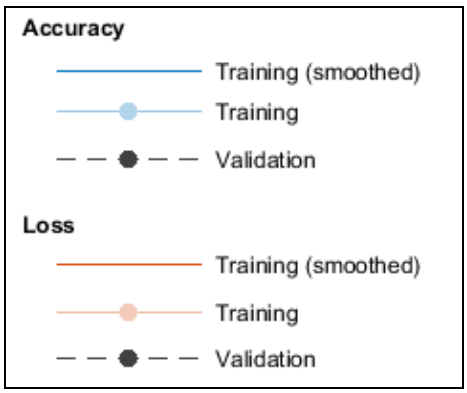

Figura 5.2: Leyenda de colores para el entrenamiento de la red

- Pruebas red Alexnet

La primera prueba se realizó con el color de pelo. En la Figura 5.3, se puede observar la gráfica superior que mide el rendimiento por época de la red Alexnet, como inicialmente se fijó el número de épocas en 30.

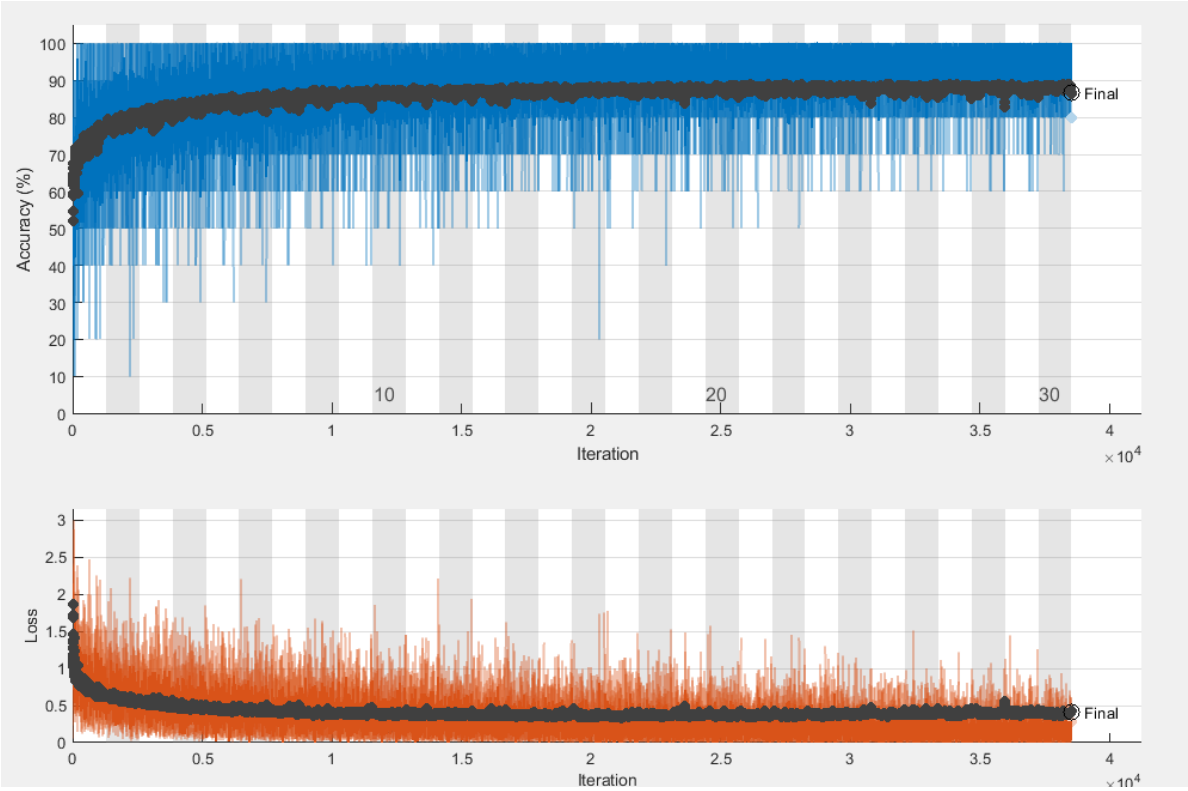

Figura 5.3: Entrenamiento red Alexnet atributo color de pelo, 30 épocas.

La red no sufre un sobreajuste debido a las técnicas que tiene para entrenar: aumento de datos y capas *dropout*, pero tampoco mejora desde la época 10. He de destacar que para entrenar la red y obtener esta simulación se tardó 70 horas de ejecución.

La gráfica inferior de la misma figura, Figura 5.3, representa las pérdidas o función de coste de este entrenamiento. Se obtiene un valor bajo, pero para lo mismo que con el rendimiento, no disminuye desde la época 10.

Por ello, se volvió a analizar el mismo atributo, pero acotando las épocas a 10. Este es la configuración fijada para este atributo.

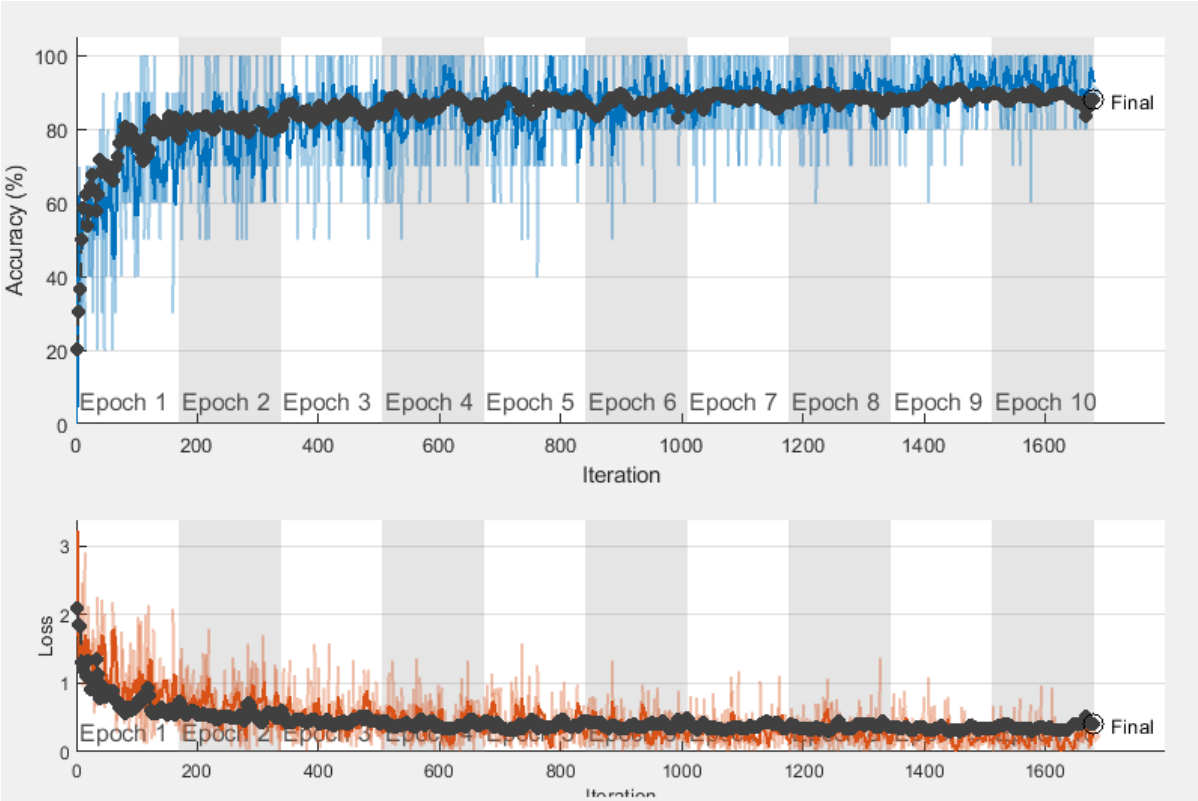

Figura 5.4: Entrenamiento red Alexnet atributo color de pelo, 10 épocas.

Tras decidir el número de épocas igual a 10 para entrenar con imágenes desbalanceadas, se debe decidir el número de épocas con imágenes balanceadas. Los datasets han sido balanceados tomando la clase con menor número de imágenes tanto en entrenamiento como validación, en este caso la clase pelo rubio:

Atributo	Clases	Número de imágenes por clase	Dataset Entrenamiento	Dataset Validación	Dataset Entrenamiento Balanceado	Dataset Validación Balanceado
Color de pelo	Pelo blanco	809	482	206	287	123
	Pelo gris	1578	939	402	287	123
	Pelo marrón	3986	2372	1016	287	123
	Pelo negro	11491	6837	2930	287	123
	Pelo rubio	482	287	123	287	123
Total		**18346**	**10917**	**4677**	**1435**	**615**

Tabla 5.4: Datasets balanceados atributo color de pelo

Se observa que el nuevo conjunto de entrenamiento es apenas un 13% del total de las imágenes utilizadas en el entrenamiento desbalanceado. Con estos datos, se realiza el entrenamiento de la red fijando el número de épocas inicialmente a 30.

Figura 5.5: Entrenamiento red Alexnet atributo color de pelo balanceado, 30 épocas.

Como se puede observar en la Figura 5.5, sí que sigue incrementando ligeramente el rendimiento con las imágenes de entrenamiento, pero se produce un sobreajuste al validar la red cada 30 iteraciones, parámetro fijado previamente. De la misma forma, se produce un sobreajuste en la función de coste.

Estos sobreajustes se producen por la falta de imágenes para entrenar la red, el equilibrio o balanceo de datos ha disminuido tanto los datasets de entrenamiento y validación que la red no es capaz de aprender correctamente. Lo que se consigue es disminuir el rendimiento de la red, pues el valor que se va a aproximar en el rendimiento con el conjunto de datos de prueba será muy similar al rendimiento obtenido con el conjunto de datos de validación.

Hay que destacar de nuevo el tiempo, pues para obtener este último rendimiento de la Figura 5.5 se ha tardado 6 minutos de ejecución. Nada comparable a las 70horas de las 30 épocas desbalanceadas previas.

- Pruebas red GoogLeNet

A continuación, se muestra los resultados obtenidos en la red GoogLeNet para el color de prenda inferior en la sección 3. Se ha realizado el mismo proceso que con la red Alexnet, fijar el número de épocas inicialmente a 30, un número alto, para observar aproximadamente en qué época deja de crecer y se estabiliza.

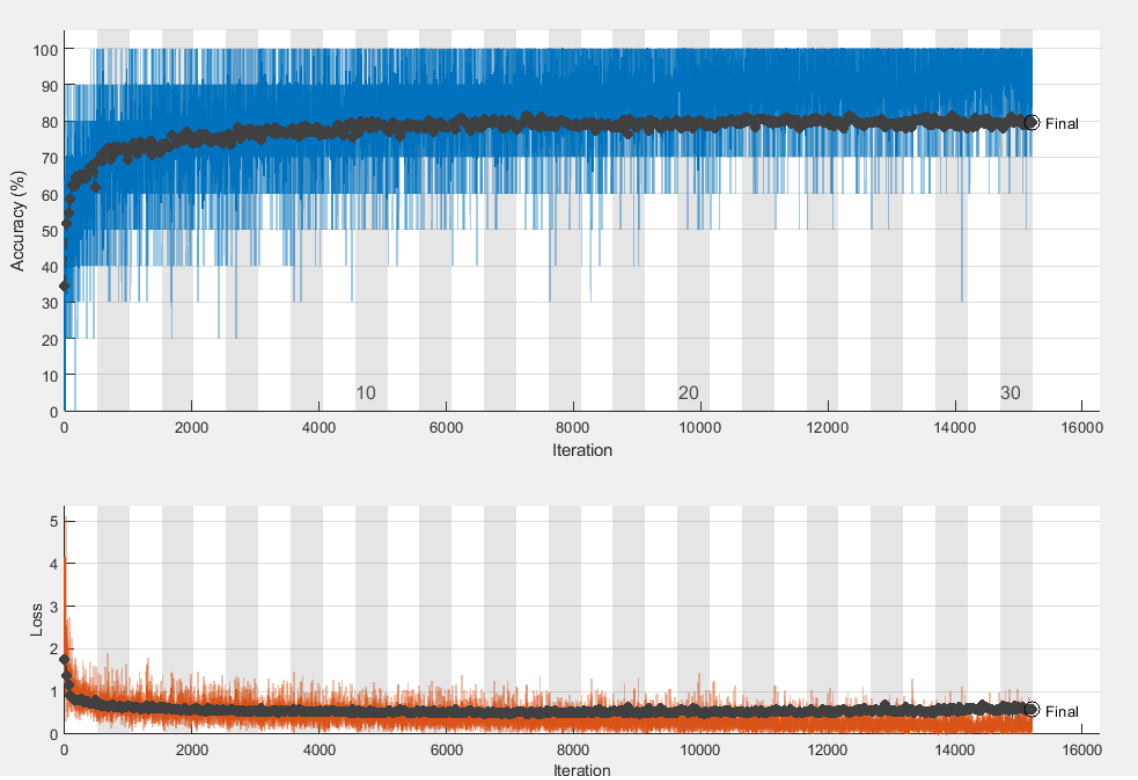

Figura 5.6: Entrenamiento red GoogLeNet atributo color de prenda inferior, 30 épocas.

Tras observar que, con un número aproximado de 10 épocas, la red puede obtener su máximo resultado, se fija estas épocas para obtener de nuevo los resultados de la red.

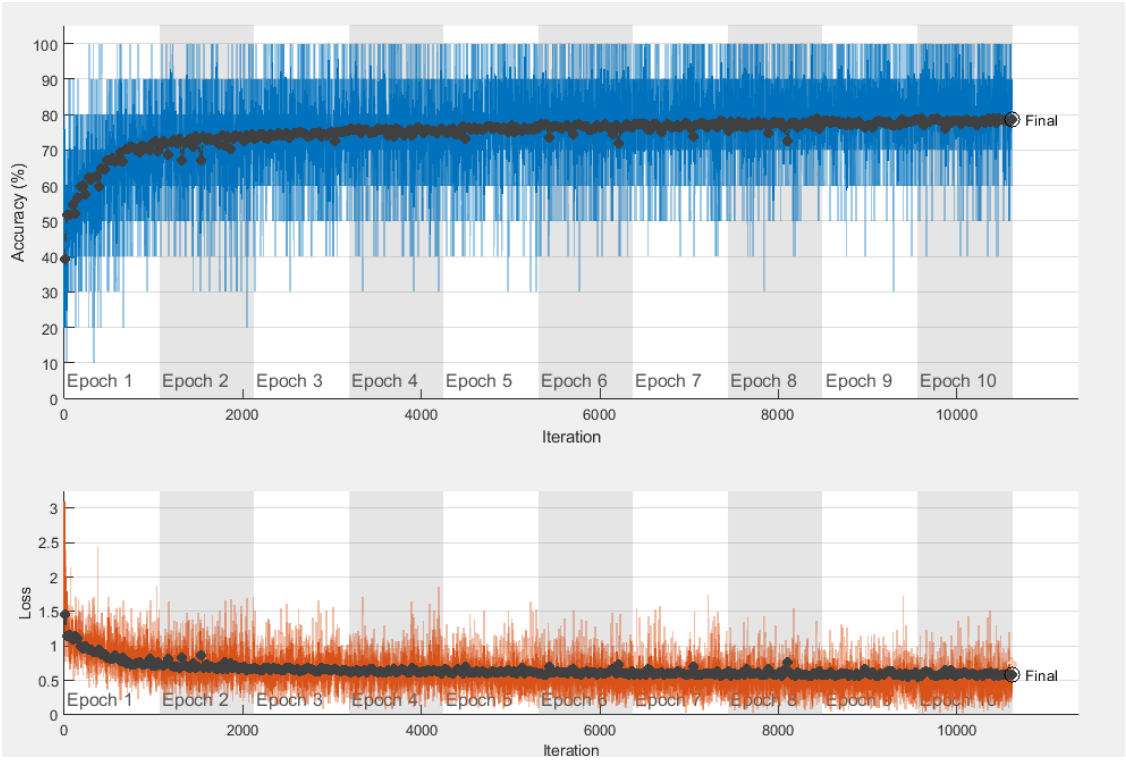

Figura 5.7: Entrenamiento red GoogLeNet atributo color de prenda inferior, 10 épocas.

Los resultados de clasificación obtenidos para este atributo son bastante buenos, pero para saber si la red no ha sido influenciada por las clases que contienen más imágenes, en este caso el color de prenda inferior negro, se vuelve a realizar el balanceado de datos.

Atributo	Clases	Número de imágenes por clase	Dataset Entrenamiento	Dataset Validación	Dataset Entrenamiento Balanceado	Dataset Validación Balanceado
Color prenda inferior	Azul	3461	2059	883	479	206
	Gris	4432	2637	1130	479	206
	Marrón	806	479	206	479	206
	Negro	9158	5449	2335	479	206
Total		**17857**	**10624**	**4554**	**1916**	**824**

Tabla 5.5: Datasets balanceados atributo color prenda inferior

Se observa que los datasets balanceados corresponden a un 18% de los datasets utilizados para entrenar este mismo atributo con clases desiguales en número de imágenes. Para estos nuevos conjuntos de datos, se vuelve a entrenar la red GoogLeNet fijando 30 épocas.

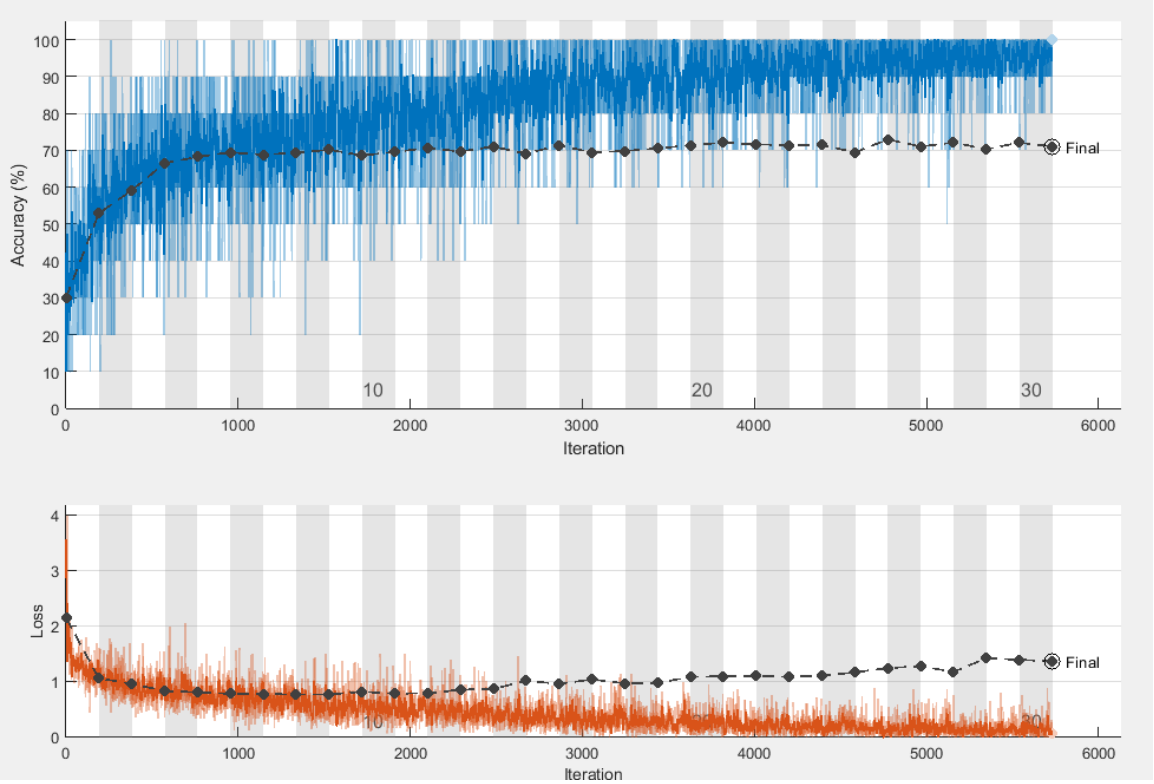

Figura 5.8: Entrenamiento red GoogLeNet atributo color de prenda inferior balanceado, 30 épocas.

Como pasaba con la red Alexnet y el atributo color de pelo, se observa un sobreajuste de la red en rendimiento y función de coste para el entrenamiento. Además, se observa que la función de coste comienda a crecer en pérdidas. Esto es debido al bajo número de imágenes que se ha utilizado para entrenar y validar al igualar las clases.

Estos mismos resultados se pueden observar tanto en la red Alexnet como GoogLeNet con otros atributos como edad o longitud de pelo, dado que tienen una de sus clases muy desigualada en número de imágenes, y cuando se balancean los datasets, disminuye en gran proporción el total de imágenes para entrenar y validar.

- Pruebas red propia

Para la red propia creada desde cero, se realizaron las mismas pruebas: fijar un número de épocas inicial y tras ver los primeros resultados ajustar el parámetro épocas al óptimo para esta red evaluando el conjunto de atributos, no uno solo.

A continuación, se muestra el entrenamiento realizado para la sección 1 del atributo color de la prenda superior de ropa con imágenes desbalanceadas y con número de épocas=30.

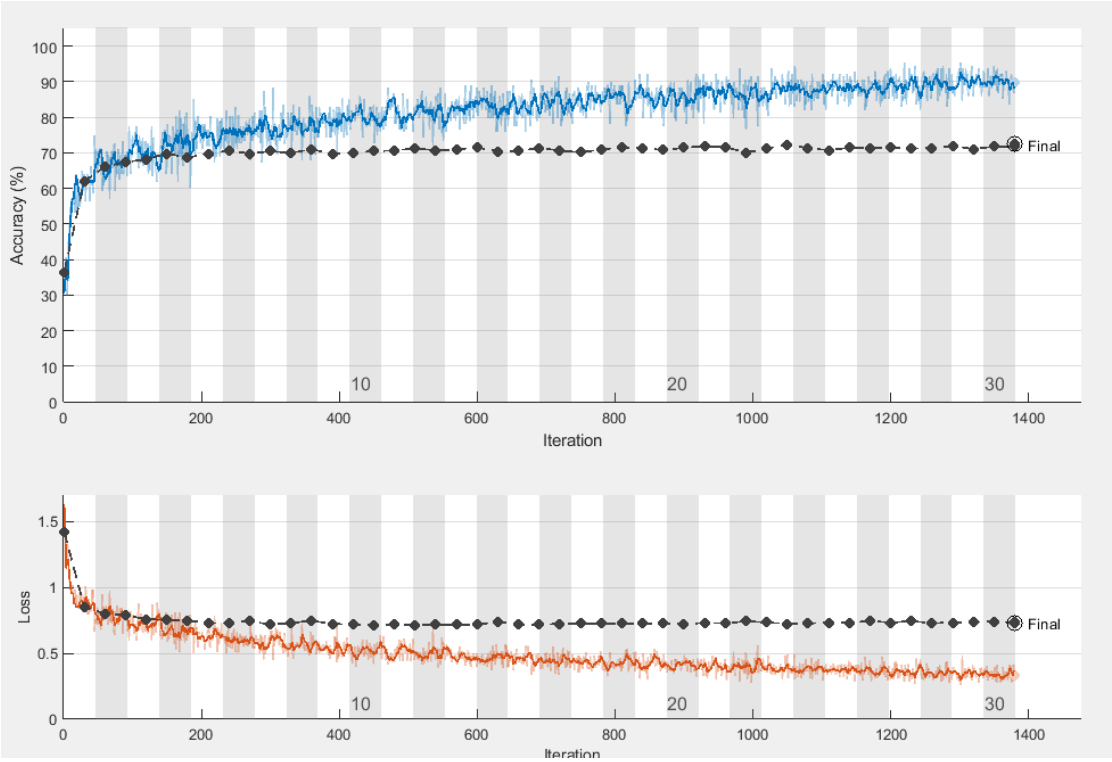

Figura 5.9: Entrenamiento red propia atributo color de prenda superior sección 1, 30 épocas.

Como se puede observar en la figura, se produce un sobreajuste importante tanto en el rendimiento de la red como en la función de coste. El rendimiento obtenido para el entrenamiento es 72,42%.

Dado que el número de épocas es alto, el número de imágenes al estar desbalanceadas es lo máximo que se puede obtener y no es un atributo que tenga unas clases tan desiguales como puede ser color de pelo, se aplicaron algunos cambios a la red.

Para evitar un sobreajuste de la red, se puede realizar tres tareas:

- Aumentar el número de imágenes para los dataset de entrenamiento y validación.
- Realizar la técnica aumento de datos: aplicando un conjunto de rotaciones y traslaciones en los pixeles de las imágenes.
- Aplicar a la red al menos una capa *dropout*, la cual para cada época del entrenamiento descarta aleatoriamente el porcentaje de neuronas fijado como parámetro.

Como cabe esperar, no se dispone de más imágenes etiquetadas para este atributo, dado que se ha usado la base de datos completa y se han reservado un conjunto de imágenes para pruebas que no se puede utilizar en el entrenamiento.

Por tanto, se ha probado a insertar una capa *dropout* con una tasa de descarte de 40% y 50% en dos entrenamientos respectivamente. Esta técnica ha descartado aleatoriamente neuronas, pero no ha conseguido mejorar el sobreajuste de la red, ni el rendimiento de esta.

Por último, se ha realizado la técnica aumento de datos con el conjunto de imágenes de entrenamiento. Se ha configurado:

- Reflexión sobre el plano X de las imágenes.
- Traslación de un rango de pixeles de [-30 30] sobre los planos X e Y.

Esta técnica sí produjo resultado, evitando por completo el sobreajuste en el entrenamiento de los atributos y ligeramente aumentó el rendimiento de la red con un resultado de 76.04%.

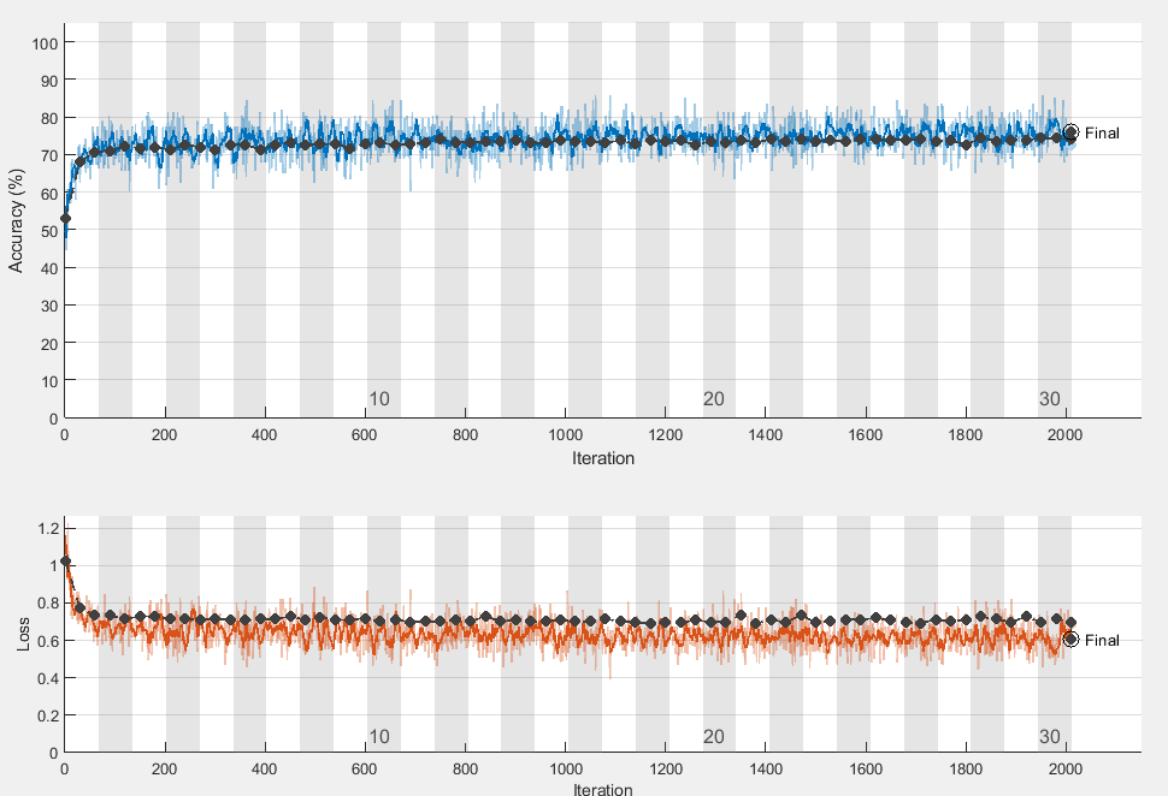

Figura 5.10: Entrenamiento red propia atributo color de prenda superior con aumento de datos, sección 1, 30 épocas.

Como estos resultados pueden verse influidos por la desigualdad de las clases, se realiza el balanceado de los conjuntos de imágenes para valorar la época óptima a la que parar el entrenamiento. Las imágenes balanceadas para este atributo quedarían de la siguiente forma:

Atributo	Clases	Número de imágenes por clase	Dataset Entrenamiento	Dataset Validación	Dataset Entrenamiento Balanceado	Dataset Validación Balanceado
Color prenda superior	Blanco	2840	1690	724	1690	724
	Gris	3130	1863	798	1690	724
	Negro	8532	5076	2176	1690	724
Total		14502	8629	3698	5070	2172

Tabla 5.6: Datasets balanceados atributo color prenda superior

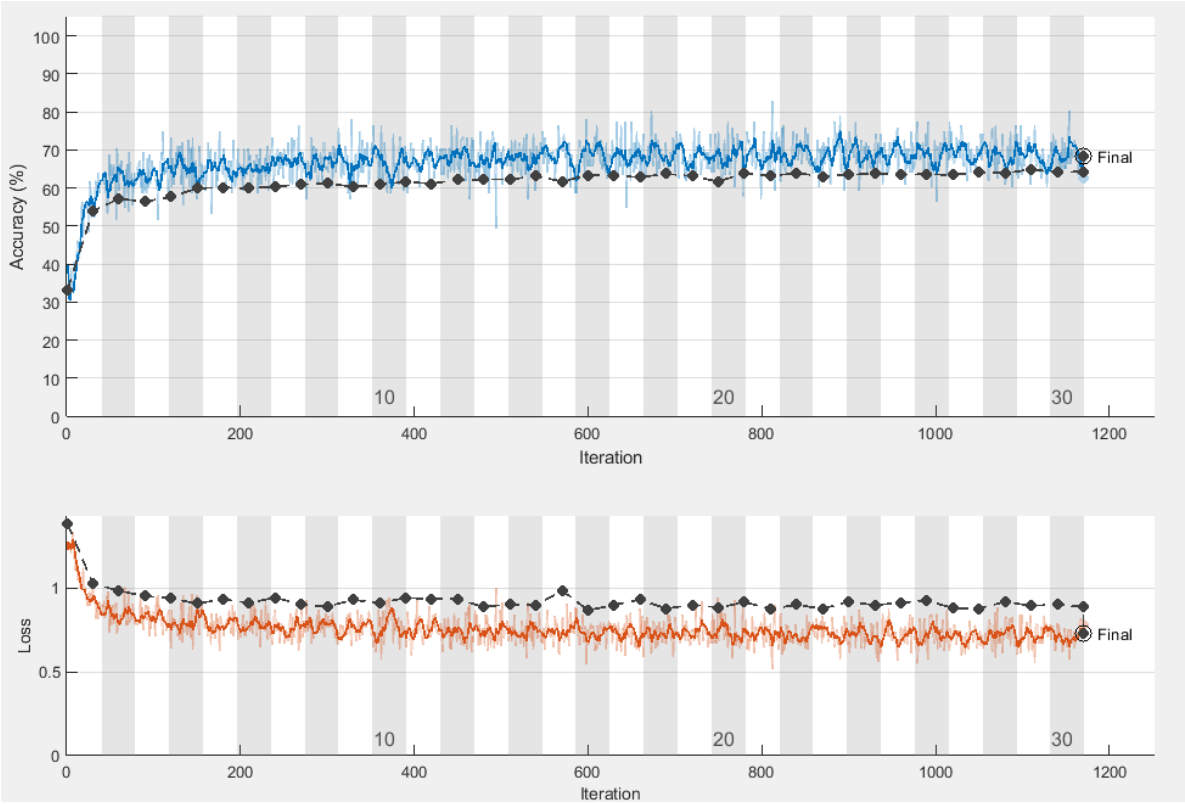

Figura 5.11: Entrenamiento red propia atributo color de prenda superior balanceado con aumento de datos, sección 1, 30 épocas.

En la Figura 5.11 se observa que existe un ligero sobreajuste en la red, de nuevo debido a la disminución de imágenes para entrenar y validar la red. El resultado del rendimiento para este entrenamiento es de 68,46%. Tras los cambios aplicados, no se podría mejorar este sobreajuste a no ser que fuese añadiendo más imágenes de cada clase.

Tras realizar diferentes pruebas con los atributos de las tres secciones, se evaluaron el conjunto de ellos para las tres redes neuronales convolucionales con los parámetros óptimos decididos.

5.3 Resultados y análisis

A continuación, se muestran los rendimientos obtenidos para cada atributo dado los conjuntos de imágenes que no han sido conocidos por los diferentes modelos antes, los conjuntos de pruebas.

En la misma tabla se va a comparar los datos tras balancear el número de imágenes en conjunto de entrenamiento y validación. El balanceado de imágenes se ha realizado con el mismo proceso realizado en las pruebas, obteniendo la clase con menor número de imágenes e igualando el resto de las clases para que tengan ese mismo número.

En primer lugar, se fija el parámetro de aprendizaje inicial de cada red decidido y tras las pruebas, el número de épocas ha sido fijado según si se apreciaba un ligero incremento a medida que entrenaba la red durante las épocas.

El conjunto de imágenes de pruebas utilizado no ha sido balanceado pues no supone alteración en el entrenamiento y así se puede tener el mismo conjunto de imágenes para comparar todas las redes y atributos.

- AlexNet
    - Tasa de aprendizaje inicial= 0.0001
    - Número de épocas imágenes no balanceadas= 10
    - Número de épocas imágenes balanceadas = 30

Como se ha demostrado en la fase de pruebas, existen 4 atributos muy desiguales: color de pelo, edad, longitud de pelo y color de la prenda inferior, este último se ve afectado en las secciones 2 y 3.

Para estos atributos se ha observado que cuando se realiza el balanceado de imágenes, se produce un sobreajuste en el entrenamiento de la red. A pesar de ello, se ha fijado el número de épocas=30 como con el resto de los atributos, porque los resultados en rendimiento son mucho mejores que si la red se entrenase solo por 6, 8 o 10 épocas, donde aparecen los puntos de inflexión comenzando el sobreajuste en estos atributos.

En la siguiente tabla se puede observar los rendimientos obtenidos para el conjunto de imágenes de pruebas para AlexNet.

Sección	Atributo	No balanceadas	Balanceadas
1	Color de pelo	84,85%	71,51%
	Edad	75,47%	50,79%
	Longitud de pelo	86,99%	59,26%
	Color prenda superior	84,05%	83,45%
	Genero	84,77%	87,02%
2	Color prenda superior	83,13%	81,84%
	Color prenda inferior	74,36%	66,74%
	Genero	77,82%	80,67%
3	Color prenda inferior	76,82%	67,53%
	Zapatos	74,18%	74,97%
	Genero	71,61%	75,89%

Tabla 5.7: Rendimiento red Alexnet por atributo

- GoogLeNet
    - Tasa de aprendizaje inicial= 0.0003
    - Número de épocas imágenes no balanceadas=10
    - Número de épocas imágenes balanceadas= 30

Para esta red, a pesar de tener ciertas mejoras en la velocidad de entrenamiento como congelar los pesos de las capas iniciales de la red, o ser una red más compleja con módulos Inception, sigue ocurriendo lo mismo que con la red Alexnet para ciertos atributos. Los mismos atributos que antes: color de pelo, edad, longitud de pelo y color de la prenda inferior en las secciones 2 y 3, sufren sobreajuste cuando las imágenes se igualan en las diferentes clases, debido al bajo número de imágenes que se dejan para entrenar y validar

Pero como los resultados se incrementan ligeramente entre la época 10 y la época 30, se ha vuelto a fijar 30 épocas para el entrenamiento con clases balanceadas, dado que el tiempo también es mucho menor y no supone tantas horas de entrenamiento como si están las clases completas.

A continuación, se muestran los resultados para las tres secciones segmentadas y el total de los atributos. El resultado es el rendimiento obtenido con el conjunto de imágenes de prueba.

Sección	Atributo	No balanceadas	Balanceadas
1	Color de pelo	88,94%	77,57%
	Edad	80,32%	64,65%
	Longitud de pelo	91,13%	79,93%
	Color prenda superior	86,63%	82,69%
	Genero	88,75%	89,66%
2	Color prenda superior	83,72%	80,82%
	Color prenda inferior	79,03%	68,03%
	Genero	81,58%	81,03%
3	Color prenda inferior	77,08%	70,06%
	Zapatos	63,20%	56,59%
	Genero	74,98%	61,64%

Tabla 5.8: Rendimiento red GoogLeNet por atributo

Como se observa, a pesar de dejar durante más épocas entrenar ambas redes, el rendimiento de clasificación disminuye considerablemente en todos los atributos respecto a los rendimientos con el conjunto de imágenes sin balancear, a excepción del atributo género en la sección 1. Este atributo va a ser analizado más adelante en profundidad.

- Red propia
    - Tasa de aprendizaje inicial= 0.00001
    - Número de épocas imágenes no balanceadas= 30
    - Número de épocas imágenes balanceadas = 30

Para la red propia como se ha explicado en las pruebas, ha sido necesario realizar la técnica de aumento de datos para generar más datos de entrenamiento a partir de las

muestras de entrenamiento existentes, “aumentando” las muestras a través de una serie de transformaciones aleatorias que producen imágenes de apariencia creíble.

El objetivo es que, en el momento del entrenamiento, el modelo nunca vea exactamente la misma imagen dos veces. Esto ayuda a que el modelo se exponga a más aspectos de los datos y se generalice mejor.

En la siguiente tabla se muestran los resultados obtenidos para el rendimiento de la red creada desde cero.

Sección	Atributo	No balanceadas	Balanceadas
1	Color de pelo	70,60%	60,39%
	Edad	58,42%	51,76%
	Longitud de pelo	73,66%	50,43%
	Color prenda superior	73,84%	70,21%
	Genero	64,67%	66,35%
2	Color prenda superior	75,22%	70,27%
	Color prenda inferior	64,35%	50,31%
	Genero	60,91%	64,65%
3	Color prenda inferior	65,06%	56,51%
	Zapatos	65,39%	61,64%
	Genero	57,51%	62,42%

Tabla 5.9: Rendimiento red propia por atributo

Resultados finales

Finalmente se van a tomar los resultados de las clases balanceadas como los resultados finales, porque a pesar de disminuir mucho el conjunto de imágenes de entrenamiento y validación, los rendimientos obtenidos para las clases desiguales no se pueden considerar válidos al poder estar influenciados por la clase con el mayor número de imágenes.

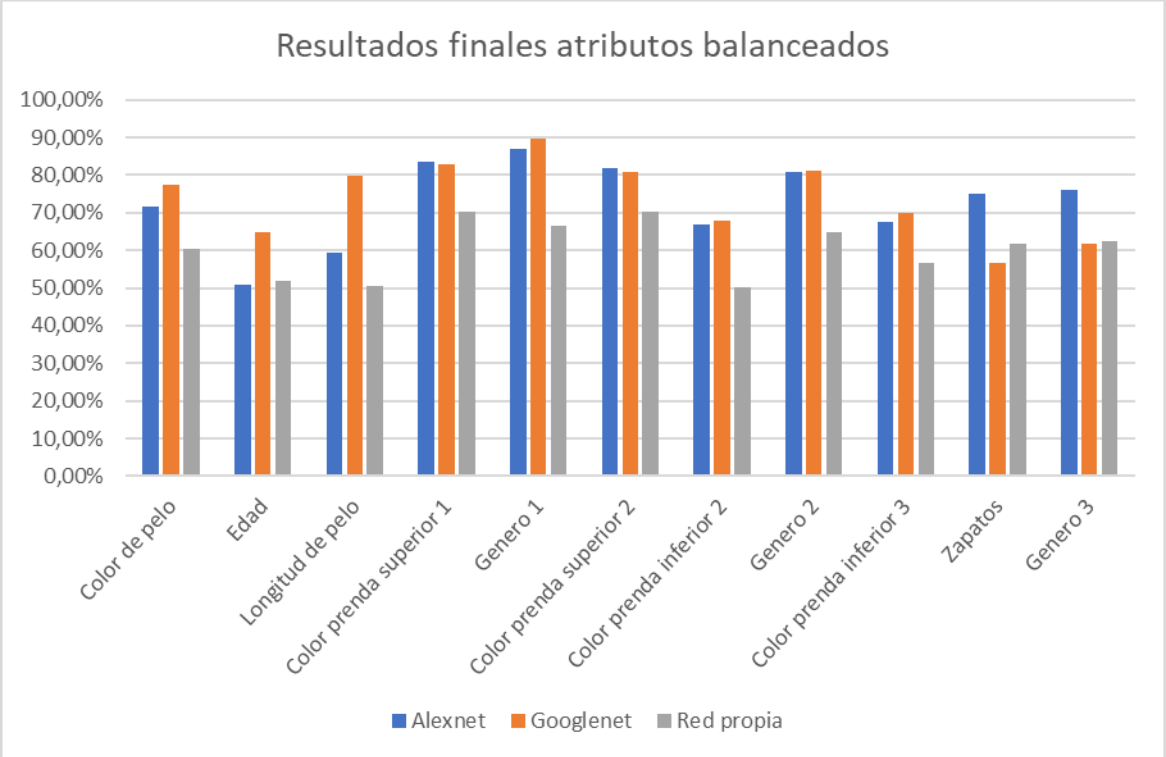

Figura 5.12: Gráfico resultados finales atributos balanceados

Los resultados obtenidos son muy satisfactorios para las tres redes al tener una media de rendimiento, a pesar de la disminución de imágenes tras el balanceado de estas, en cada red de: Alexnet=72.70%, GoogLeNet=73.87%, Red Propia= 60.44%.

Claramente, se observa que la red GoogLeNet obtiene los mejores rendimientos de clasificación para los atributos seleccionados. Esto es debido a la complejidad que tiene la red con los módulos Inception, las técnicas de aumento de datos, además del conjunto de opciones de entrenamiento para combinar las velocidades de entrenamiento en las diferentes capas.

Los módulos Inception han realizado la convolución con diferentes tamaños de filtros en paralelo, lo que ha producido que sea capaz de extraer más características de cada sección.

Entre todos los atributos, el género de los viandantes es el más destacable porque es el único que ha aumentado su rendimiento tras igualar el número de imágenes por clase. Esto es debido a que es el atributo que utiliza mayor número de imágenes en total 18997, casi el total de la base de datos PETA, y entre ellas existe el ratio más igualado de 55%-45% entre sus clases, hombres y mujeres.

En la red propia aumenta en las tres secciones segmentadas de las imágenes, y nunca baja de un rendimiento inferior al 60% un aspecto muy bueno de la red porque identificar el género del individuo por las piernas-pies es algo muy complicado hasta para una persona. Siendo un problema con dos clases, se consideraría un resultado aleatorio si la tasa estuviese alrededor del 50%, pero las tres redes y secciones superan este valor.

A continuación, se muestra una prueba final realizada con la red que ha obtenido los mejores resultados, GoogLeNet balanceado para cada atributo. Se ha probado con una imagen de la base de datos original, PETA, rotada simétricamente sobre el eje horizontal. Esta imagen corresponde a la base de datos VIPeR imagen 828, segmentada en las tres secciones y reajustada en tamaño 224x224 pixeles.

Figura 5.13: Prueba de ejemplo sobre imagen simétrica en el eje horizontal

Los resultados obtenidos se pueden ver en la tabla 5.10, donde se muestra el mayor porcentaje obtenido para cada atributo y con el nombre de esa clase.

Como se puede observar, se ha marcado en naranja los resultados erróneos por la red, confundiendo con mayor error el color de la prenda inferior en la sección 2. Además se puede ver como el rango de edad para este viandante no sería correcta hasta la tercera clase con mayor probabilidad, asociando a un rango de edad mayor antes.

Sección	Atributo	Clases	Red GoogLeNet Balanceada	Resultado correcto
1	Color de pelo	Pelo blanco	0,04%	**Pelo negro**
		Pelo gris	0,00%	
		Pelo marrón	0,00%	
		Pelo negro	**99,96%**	
		Pelo rubio	0,00%	
	Edad	Menor de 15 años	0,32%	**Entre 15 y 30**
		Comprendida entre 15 y 30 años	1,55%	
		Comprendida entre 30 y 45 años	**65,66%**	
		Comprendida entre 45 y 60 años	32,43%	
		Mayor de 60 años	0,04%	
	Longitud de pelo	Pelo largo	4,86%	**Corto**
		Pelo corto	**95,12%**	
		Calvo	0,02%	
	Color prenda superior	Blanco	**99,98%**	**Blanco**
		Gris	0,02%	
		Negro	0,00%	
	Género	Hombre	**99,80%**	**Hombre**
		Mujer	0,20%	
2	Color prenda superior	Blanco	**85,12%**	**Blanco**
		Gris	10,33%	
		Negro	4,55%	
	Color prenda inferior	Azul	0,04%	**Marrón**
		Gris	**95,20%**	
		Marrón	0,75%	
		Negro	4,02%	
	Género	Hombre	43,30%	**Hombre**
		Mujer	**56,70%**	
3	Color prenda inferior	Azul	0,00%	**Marrón**
		Gris	3,19%	
		Marrón	**96,81%**	
		Negro	0,00%	
	Zapatos	Zapatillas	**57,44%**	**Zapatillas**
		Zapatos	42,56%	
	Genero	Hombre	**68,88%**	**Hombre**
		Mujer	31,12%	

Tabla 5.10: Resultados prueba de ejemplo con imagen rotada simétricamente en eje horizontal

6. Conclusiones y trabajo futuro

En este capítulo se va a realizar una valoración global del trabajo que se ha desarrollado. En primer lugar, se realizará una valoración personal para ver si se han cumplido los objetivos que fueron marcados al inicio del proyecto, a continuación, se detallará el aprendizaje conseguido y finalmente se propondrán algunas vías de desarrollo futuras.

6.1 Conclusiones

El principal objetivo marcado al inicio del proyecto fue principalmente conocer en profundidad una red convolucional neuronal: el funcionamiento, sus capas y sus métodos para la clasificación de atributos con el fin de implementar un sistema de clasificación de atributos de personas en la distancia.

Hasta ahora, la clasificación de rasgos biométricos suaves mediante sistemas de aprendizaje profundo no ha conseguido obtener unos resultados mejores que la clasificación manual de estos. Los sistemas se ven influidos por la calidad de las imágenes, así como sombras, ángulos, distancia u ocultación de atributos por otras personas o cosas del entorno.

Los rasgos biométricos suaves, pueden aportar mucha información para los sistemas de reconocimiento o identificación, pero al no ser rasgos únicos y exclusivos de cada persona, no sirven para un sistema fiable con alta seguridad.

Los resultados obtenidos para las redes previamente entrenadas Alexnet y Googlenet son satisfactorios dado que estas redes no han sido entrenadas previamente para este tipo de atributos, y, además, no toman como relevante el color en sus clasificaciones. Se ha obtenido una tasa de acierto muy elevada.

Pero lo más destacable son los resultados obtenidos en la red propia. Esta red ha sido generada totalmente desde el inicio con sus capas configurando los parámetros y las opciones de entrenamiento y los resultados se aproximan en muchos atributos a las redes previamente entrenadas, que son redes mucho más complejas y galardonadas.

En la gran mayoría de atributos, los resultados en rendimiento podrían ser superiores si se tuviese más imágenes para entrenar la red, dado que se ha utilizado una única base de datos con 19mil imágenes, un número muy por debajo de lo que se usa para estos casos. Véase el ejemplo de Alexnet o GoogLeNet que han sido entrenadas por más de un millón de imágenes.

Por este motivo y dado que las redes inteligentes todavía no superan el conocimiento de las personas, los resultados obtenidos en las tres redes han sido muy satisfactorios.

Para finalizar, resaltar que se han cumplido los objetivos marcados al inicio de este Trabajo Fin de Máster, se ha ampliado el conocimiento aprendido en la asignatura Biometría del Máster realizado, a cerca de la extracción y clasificación de rasgos biométricos suaves, así como del aprendizaje profundo mediante el uso de redes neuronales convolucionales.

6.2 Trabajo futuro

Como posibles vías de trabajo futuro se tienen los siguientes puntos:

- Aumentar la base de datos etiquetada de atributos en peatones, con el fin de mejorar el entrenamiento y clasificación de las CNN y que conozcan más situaciones de viandantes con los mismos e incluso otros atributos. Se propone añadir el atributo “teléfono móvil”, indicando si el viandante se encuentra hablando por teléfono o manejando su terminal móvil.
- Perfeccionar la red neuronal convolucional propia, mediante la adición de capas a la red, si fuese necesario, y ajuste de parámetros para obtener un rendimiento superior.
- Entrenar la red para imágenes de cuerpo entero en la distancia.
- Aplicar las redes más optimas en lugares de interior como universidades o aeropuertos para mejorar la seguridad de estos.

Referencias

[1]Paloma Recuerdo de los Santos 16 noviembre 2017. [En línea]. Available: <<https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/,>> [Último acceso: 5 agosto 2020].

[2]F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, J.Xiao, “LSUN: Construction of a Large-Scale Image Dataset using Deep Learning with Humans in the Loop,” Jun 2016

[3]N. Singh Chauhan, 3 octubre 2019. [En línea]. Available: <<https://towardsdatascience.com/introduction-to-artificial-neural-networks-ann-1aea15775ef9,>> [Último acceso: 14 agosto 2020]

[4]Na8 12 septiembre 2018. [En línea]. Available: <<https://www.aprendemachinelearning.com/breve-historia-de-las-redes-neuronales-artificiales/,>> [Último acceso: 5 agosto 2020].

[5]D. Svozil, V.Kvasnicka, J.Pospíchal, “Introduction to multi-layer feed-forward neural networks” Chemometrics and Intelligent Laboratory Systems 39 (1997) 43-62. PZZ SO169-7439(97)00061-O

[6]Dipanjan Sarkar, 14 noviembre 2018. [En línea]. Available: <<https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a,>> [Último acceso: 13 agosto 2020]

[7]F. Radenovic, G. Tolias, O. Chum, “Fine-tuning CNN Image Retrieval with No Human Annotation,” julio 2018, arXiv:1711.02512v2.

[8]M. Sultana, P. P. Paul and M. Gavrilova, "A Concept of Social Behavioral Biometrics: Motivation, Current Developments, and Future Trends," 2014 International Conference on Cyberworlds, Santander, 2014, pp. 271-278, doi: 10.1109/CW.2014.44.

[9]Al-Maadeed et al. EURASIP Journal on Image and Video Processing (2016) 2016:1 DOI 10.1186/s13640-015-0097-y

[10] R. P. Krish, J. Fierrez, D. Ramos, J. Ortega-Garcia and J. Bigun, "Pre-Registration of Latent Fingerprints based on Orientation Field", IET Biometrics, Vol. 4, pp. 42-52, June 2015.

[11] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez and F. Alonso-Fernandez, "Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation", IEEE Trans. on Information Forensics and Security, Vol. 13, n. 7, 2018.

[12] P. Tome, R. Vera-Rodriguez, J. Fierrez and J. Ortega-Garcia, "Facial Soft Biometric Features for Forensic Face Recognition", Forensic Science International, Vol. 257, pp. 171-284, December 2015.

[13] R. Vera-Rodriguez, M. Blazquez, A. Morales, E. Gonzalez-Sosa, J. Neves and H. Proenca, "FaceGenderID: Exploiting Gender Information in DCNNs Face Recognition Systems", in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Workshop on Bias Estimation in Face Analytics (CVPR BEFA), June 2019 (Best Paper Runner Up Award).

[14] R. Tolosana, R. Vera-Rodriguez, J. Fierrez and J. Ortega-Garcia, "DeepSign: Deep On-Line Signature Verification", arXiv preprint arXiv:2002.10119, 2020.

[15] R. Vera-Rodriguez, R. Tolosana, J. Hernandez-Ortega, A. Acien, A. Morales, J. Fierrez and J. Ortega-Garcia, "*Modeling the Complexity of Signature and Touch-Screen Biometrics using the Lognormality Principle*", Rejean Plamondon and Angelo Marcelli and Miguel A. Ferrer (Eds.), The Lognormality Principle and its Applications, World Scientific, 2020.

[16] A. Acien, J. V. Monaco, A. Morales, R. Vera-Rodriguez and J. Fierrez, "TypeNet: Scaling up Keystroke Biometrics", in IEEE/IAPR Intl. Joint Conf. on Biometrics (IJCB), September 2020

[17] J. Franco-Pedroso and J. Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition", Speech Communication, Elsevier Science Publishers B. V., Vol. 76, pp. 61-81, February 2016.

[18] O. C. Reyes, R. Vera-Rodriguez, P. Scully and K. B. Ozanyan, "Analysis of Spatio-temporal Representations for Robust Footstep Recognition with Deep Residual Neural Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, n. 99, 2018.

[19] R. Vera-Rodriguez, J. S. Mason, J. Fierrez and J. Ortega-Garcia, "Comparative Analysis and Fusion of Spatio-Temporal Information for Footstep Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence , Vol. 35, n. 4, pp. 823-834, April 2013.

[20] R. Vera-Rodriguez, N. W. D. Evans and J. S. D. Mason, "Footstep Recognition", Stan Z. Li and Anil K. Jain (Eds.), Encyclopedia of Biometrics, Springer, pp. 693-700, 2015 (ISBN 978-1-4899-7487-7, re-edited from 2009).

[21] R. Tolosana, R. Vera-Rodriguez, J. Fierrez and J. Ortega-Garcia, "BioTouchPass2: Touchscreen Password Biometrics Using Time-Aligned Recurrent Neural Networks", IEEE Transactions on Information Forensics and Security, 2020.

[22] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome and J. Fierrez, "Measuring the Gender and Ethnicity Bias in Deep Models for Face Recognition", in Proc. of IAPR Iberoamerican Congress on Pattern Recognition, CIARP, Springer, pp. 584-593, Madrid, Spain, November 2018.

[23] P. Tome, J. Fierrez, R. Vera-Rodriguez and M. Nixon, "Soft Biometrics and their Application in Person Recognition at a Distance", IEEE Transactions on Information Forensics and Security, Vol. 9, n. 3, pp. 464-475, March 2014.

[24] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Z. Li, “Cross dataset person reidentification,” in Proc. ACCV Workshop, 2014

[25] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, “Learning to rank in person re-identification with metric ensembles,” in Proc. CVPR, 2015

[26] S. Wu, Y.-C. Chen, and W.-S. Zheng, “An enhanced deep feature representation for person re-identification,” in Proc. WACV, 2016.

[27] D. i, X. Chen, and K. Huang, “Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios,” in Proc. ACPR, 2015.

[28] J. hu, S. Liao, D. Yi, Z. Lei, and S. Z. Li, “Multi-label CNN based pedestrian attribute learning for soft biometrics,” in Proc. ICB, 2015.

[29] P. udowe, H. Spitzer, and B. Leibe, “Person attribute recognition with a jointly-trained holistic cnn model,” in Proc. ICCV, 2015.

[30] Y. Deng, P. Luo, C. C. Loy, X. Tang, "Pedestrian attribute recognition at far distance," in Proceedings of ACM Multimedia (ACM MM), 2014

[31] R. Layne, T. M. Hospedales, S. Gong, et al. Person re-identification by attributes. BMVC, 2012.

[32] T. Nortcliffe. People analysis cctv investigator handbook. Home Office Centre of Applied Science and Technology, 2011

[33] P. Viola and M. Jones. Object detection framework. 2001

[34] Georghiades, A.S. and Belhumeur, P.N. and Kriegman, D.J. From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. IEEE Trans. Pattern Anal. Mach. Intelligence 23(6):643-660 (2001).

[35] Seely, Richard & Samangooei, Sina & Middleton, Lee & Carter, John & Nixon, Mark. (2008). The University of Southampton Multi-Biometric Tunnel and introducing a novel 3D gait dataset. 1 - 6. 10.1109/BTAS.2008.4699353.

[36] Base de datos Pedestrian Attribute. Available: https://www.dropbox.com/s/52ylx522hwbdxz6/PETA.zip?dl=0

[37] A. Krizhevsky, I.Sutskever, G.E.Hinton, “Imagenet Classification with Deep Convolutional Neural Networks”. ImageNet LSVRC-2010.

[38] C.Szegedy, W. Liu, Y. Jia, P. Sermanet, S.Reed, D.Anguelov, D. Erhan, V.Vanhoucke, A. Rabinovich, “Going Deeper with Covolutions”. ImageNet LSVRC-2014.